Big data and Official Statistics: some evidences

Big Data and Official Statistics: some evidences
Big Data e Statistiche Ufficiali: alcune evidenze

Paolo Righi, Natalia Golini, Gianpiero Bianchi

Abstract The paper compares two classes of estimators exploiting a Big Data source. Both classes rely on a probabilistic sampling. Nevertheless, while the first class of estimators uses the Big Data as auxiliary information, the latter uses the probabilistic sample as auxiliary information. We denote this second class as pseudo-calibration estimators, since it applies the calibration to a not random sample. We present an original application of the jackknife method for the variance estimation for the pseudo calibration estimators. Finally, an empirical evaluation on a real survey and Big Data compares several estimators of the two classes with a standard design-based survey estimator.

Abstract *Il lavoro confronta due classi di stimatori che sfruttano una fonte Big Data. Entrambe le classi si affidano su un campione probabilistico. Ma, mentre la prima classe usa i Big Data come informazione ausiliaria, la seconda classe usa il campione probabilistico come informazione ausiliaria. Denotiamo la seconda classe come stimatori di pseudo- calibrazione, poiché ai applica la calibrazione a un campione non casuale. Si presenta una applicazione originale del metodo jackknife per la stima della varianza per gli stimatori pseudo-calibrati. Infine, si confrontano empiricamente su dati di indagine e di Big Data reali alcuni stimatori delle due classi con uno stimatore standard design-based.*

Key words: Calibration, Big Data, Official Statistics.

1 Informative context and notation

New sources of data have emerged and are the result of more and more interactions with digital technologies by citizens and business units and the

[1] Paolo Righi, Istat; e-mail: parighi@istat.it

[2] Natalia Golini, Università degli Studi di Torino; e-mail: natalia.golini@unito.it

[3] Gianpiero Bianchi, Istat; e-mail: gianbianchi@istat.it:

increasing capability of these technologies to provide digital trails. These sources commonly referred to as Big Data, offer new challenges from the statistical viewpoint in particular generated by a paradigm shift: from designed data for planned statistics to data-oriented or data-driven statistics. Beyond the descriptive statistics, it is necessary to determine under which conditions make valid inference using Big Data. The aim to produce Official Statistics with high-quality standards has stimulated the definition of suitable statistical frameworks (among others: Eurostat, 2018; the American Association for Public Opinion Research (AAPOR) task Force on Big Data, 2015) and quality frameworks (UNECE, 2014).

The paper compares two classes of estimators that use the Big Data source for producing Official Statistics. The two sets of estimators rely on a probabilistic survey but different approach to the inference. The former class concerns a design-based framework, while the latter a model-based framework although an automatic calibration procedure, typical of a model-assisted estimator is carried out. Both classes of estimators apply the calibration techniques and make the estimators appealing to the National Statistical Institutes (NSIs), being these techniques well known by the NSIs. Section 2 introduces the basic notation and the informative context. Section 3 shows the first class of estimators, denoted as data integration estimators (Kim and Tam, 2021). Section 4 illustrates the second class of estimators, referred to this paper as pseudo-calibrated estimator (Righi *et all.*, 2021) or calibration adjustment (Lee and Valliant, 2009). Section 5 shows an empirical evaluation on real survey and Big Data. Finally, Section 6 gives some conclusions.

2 Informative context and notation

Let U be the target population of size N and $U_B \subset U$ be the sub-population of size N_B.

We denote with U_B a Big Data source. In U_B are collected or predicted by a statistical model (with a model error) the random variable $\boldsymbol{y}$. Let us denote with y_k the collected value on the unit $k \in U_B$ and with $\tilde{y}_k$ the predicted value. We use y_k^* notation to indicate either y_k or $\tilde{y}_k$. In case of more than one variable collected or predicted in the Big Data source, we have the $\mathbf{y}_k^* = (y_k^{1*}, \dots, y_k^{h*} \dots, y_k^{H*})'$ vector, being y_k^{h*} the values of hth variable collected or predicted in the Big Data. Furthermore, let $U_{\bar{B}}$ be the set of units without information from the Big Data source being $U_B \cup U_{\bar{B}} = U$ and $U_B \cap U_{\bar{B}} = \emptyset$. Let δ_k indicate the Big Data membership variable, with $\delta_k = 1$ when $k \in U_B$ and $\delta_k = 0$ when $\in U_{\bar{B}}$. Along with U_B, let s be the reference survey sample, in which a probabilistic sample is drawn from U. This is a multi-purpose survey collecting $\mathbf{y}_k$ and a vector $\mathbf{z}_k = (z_k^1, \dots, z_k^q, \dots, z_k^M)'$ of M variables for each $k \in s$. In this setting we assume to know the value of δ_k and we can define $s = s_B \cup s_{\bar{B}}$ with $s_B \cap s_{\bar{B}} = \emptyset$, with $s_B \subset U_B$ and $s_{\bar{B}} \subset U_{\bar{B}}$. Unit nonresponses could affect the reference survey sample. We indicate with r the sample of respondents.

Finally, let $\mathbf{x}_k = (x_k^1, \dots, x_k^p, \dots, x_k^P)'$ be the value vector of the P auxiliary variables known for each $k \in U$. The target parameter is the total

$$Y = \sum_U y_k \,. \tag{2.1}$$

We also consider the total for the domain $U_{(d)} \subset U$ $(d = 1, \dots, D)$,

$$Y_{(d)} = \sum_U y_k \, \lambda_{k(d)}, \tag{2.2}$$

with $\lambda_{k(d)} = 1$ if $k \in U_{(d)}$ and $\lambda_{k(d)} = 0$ otherwise. We indicate with $\boldsymbol{\lambda}_k = (\lambda_{k(1)}, \dots, \lambda_{k(d)}, \dots \lambda_{k(D)})'$ the domain indicator variable vector.

3 Data Integration estimators

We compare two classes of estimators that use in a different way the information coming from the Big Data source. We refer to the first class of estimators as Data Integration (DI) estimators (Kim and Tam, 2021). These estimators define a general tool for making proper use of the Big Data sources for finite population inference by combining the sources with a probabilistic survey.

The DI estimators are design-based, and the Big Data variables are used as auxiliary variables. It is worthy to note that the making design-based inference is an appealing property for the NSIs that usually apply this kind of approach for data production of Official Statistics. The general form of the DI estimators is the Regression DI (RegDI) estimator. By specifying the terms of the RegDI estimators, we can obtain the different DI estimators. Therefore, we focus on the general RegDI estimator. Kim and Tam (2021) give insight on the specific estimators.

The standard survey regression adjusts the survey weights to respect some known totals. In particular the following optimization problem is performed

$$\begin{cases} min \ \sum_s Q(d_k, w_k) \\ \sum_s \mathbf{x}_k \, w_k = \mathbf{X} \end{cases}, \tag{3.1}$$

where d_k is the base sampling weight, w_k is the weight of calibration, $\mathbf{X} = \sum_U \mathbf{x}_k$ is a vector of totals, that we assume as known or estimated by a large and accurate survey (e.g., Dever and Valliant, 2010, 2016) with $\mathbf{x}_k$ known for each $k \in s$.
and,

$$Q(d_k, w_k) = \sum_s d_k (\frac{w_k}{d_k} - 1)^2.$$

The RegDI estimator augments the number of auxiliary variables with δ_k and $\delta_k y_k^*$. The estimator is

$$\hat{Y}_{RegDI} = \sum_s y_k w_k \tag{3.2}$$

with $\sum_s \delta_k w_k = N_B$ and $\sum_s \delta_k y_k^* = \sum_{U_B} y_k^*$.

The domain estimator is given by

$$\hat{Y}_{RegDI(d)} = \sum_s \delta_k y_k^* w_k \lambda_{k(d)} \tag{3.3}$$

REMARK 4.1: Kim and Tam (2021) deal with the case of unknown δ_k for $k \in s$. We do not analyse this setting in this work.
REMARK 4.2: The $\hat{Y}_{RegDI}$ (3.2)-(3.3) is variable-specific. A more general expression can calibrate the weights on the auxiliary vector $(\mathbf{x}_k, \delta_k, \delta_k \mathbf{y}_k^*)$.

4 Pseudo-calibration estimators

The second class of estimators uses the Big Data source as the large non-probability sample. A critical issue when using a non-probability sample is to dealt with the unknown sample selection mechanism. In particular, since $U_B \subset U$ the no data observations in $U_{\bar{B}}$ (missing data) weakens the representativeness of the Big Data sample with respect to the target population. According to Buelens *et al*. (2014), representativeness is defined as follows: "A subset of a finite population is said to be representative of that population with respect to some variable, if the distribution of that variable within the subset is the same as in the population. A subset that is not representative is referred to as selective." Meng (2018) underlines that among the different terms generating the selection bias the most important is the correlation between $\boldsymbol{y}$ and δ. When the variable are not correlated, we do not have selection bias. In presence of correlation, there are several approaches for adjusting the selection bias in Big Data. For instance, Kim and Wang (2019), Chen *et all.* (2020), Elliot and Valliant (2017). Here we focus on the estimation process denoted as calibration weighting (Kim, 2022), calibration adjustment (Lee and Valliant, 2009) or pseudo-calibration estimator (Righi *et all.*, 2019).

The estimator calibrates the Big Data distributions on the auxiliary variables related to the target variable so that after this step, these distributions are coherent with the distributions on the target population.

To achieve this objective the calibration process assigns to each unit of the Big Data a final weight acting to satisfy the calibration constraints.

The final weights are obtained by the solution of the following optimization problem:

$$\begin{cases} min \sum_{U_B} d(p_k, w_k) \\ \sum_{U_B} \mathbf{x}_k w_k = \mathbf{X} \end{cases}, \tag{4.1}$$

where $d(.)$ is a convex function, denoted as a distance function, (Deville and Särndal, 1992), p_k is the initial weight, w_k is the weight of calibration, $\mathbf{X} = \sum_U \mathbf{x}_k$ is a vector of totals, that we assume as known or estimated by a large and accurate survey (e.g., Dever and Valliant, 2010, 2016) with $\mathbf{x}_k$ known for each $k \in U_B$.

We can fix the p_k values in different ways. If we perform a propensity adjustment (Kim, 2022, Elliot and Valliant, 2017, Lee and Valliant, 2009), p_k is the propensity of each unit to be included in the Big Data source. A statistical model estimates this propensity.

In the simplest form $p_k = N/N_B$. When $p_k = p, \forall\, k \in U_B$ and varying p the solution of the optimization problem does not change. So that, with $p_k = 1$ or $p_k = N/N_B$ the calibrated weights, w_k, are the same.

Considering $p_k = 1$ we define a statistical framework where U is a take-all sample (census) with $pr(\delta_k = 1) = 1$ for $\forall\, k \in U$. Nevertheless, U is affected by a kind of unit non-response (alternatively U_B under-covers U). The inclusion probabilities of the respondents, the units in U_B, are adjusted for reducing nonresponse bias by a calibration approach (Little and Rubin, 2007).

4.1 Observe the target variable in the Big Data source

When we collect y_k for $\in U_B$, the pseudo-calibrated estimator is given by

$$\hat{Y}_{PC,B} = \sum_U \delta_k y_k w_k \tag{4.2}$$

being $w_k = 0$ when $\delta_k = 0$. We apply the calibration algorithm (Deville and Särndal, 1992) to solve the optimization problem in (4.1).

The domain estimator is $\lambda_{k(d)}$

$$\hat{Y}_{PC,B(d)} = \sum_U \delta_k y_k w_k \lambda_{k(d)} \tag{4.3}$$

Further discussion on the $\hat{Y}_{PC,B}$ estimator is given in Righi *et all.* (2019).

REMARK 5.1: The proposed estimator has a simple and straight implementation. It leverages well-known and widely used statistical calibration tools in the NSIs.

REMARK 5.2: The proposed estimator is model-based. However, it applies the same process for adjusting unit non-response. The calibration will be generally based on a usual set of auxiliary variables exploited for calibrating or adjusting for non-response the probabilistic sample. The similarity of the process facilitates the consistency of the estimates on the same target population and the estimation computed using either the Big Data or a standard survey based on a probabilistic sample.

4.2 Predict the target variable in the Big Data source

The $\hat{Y}_{PC,B}$ is applicable when we collect y_k in U_B. In some case, we have from a Big Data source a prediction of $\tilde{y}_k$. An example of predicted data is the remote sensing for agricultural statistics (on land use, crop type, crop yield) using the satellite imageries. Another example of predicted data comes from business statistics on the services and functionalities of the enterprise's websites. To count the websites offering specific services (e-commerce, link to social media, job advertisement) we can apply a web-scraping technique by collecting text documents on the website and predicting the presence of the functionalities and services in the website by performing a text analysis and classification by machine learning techniques.

In this case, the estimator (4.2) or (4.3) has to be refined plugging-in the $\tilde{y}_k$ synthetic values for y_k,

$$\hat{Y}^{P}_{PC,B} = \sum_U \delta_k \tilde{y}_k w_k, \tag{4.4}$$

where $\tilde{y}_k$ is null for $\delta_k = 0$. The estimator for the domain $U_{(d)}$ adds the terms $\lambda_{k(d)}$ in the (4.4).

The estimator (4.4) assumes the form of the *projection estimator*. Kim and Rao (2012) define a model-assisted framework of the estimator (4.4) with $\tilde{y}_k = \xi(\mathbf{a}_k \hat{\boldsymbol{\gamma}})$ being ξ a known function, $\mathbf{a}_k$ a vector of auxiliary variable known for $k \in U$ and the $\hat{\boldsymbol{\gamma}}$ vector the estimate of the model parameter vector obtained from a second survey (the reference survey) using the data set $\{(y_k, \mathbf{a}_k): k \in s \subset U\}$ and the survey weights. Kim and Rao (2012) define the conditions to have unbiased estimates. When the conditions are not satisfied, an unbiased estimator is

$$\hat{Y}^{D}_{PC,B} = \hat{Y}^{P}_{PC,B} + \sum_{s \subset U} \delta_k (y_k - \tilde{y}_k) f_k, \tag{4.5}$$

in which the second term of the right-hand side of the (4.5) is the bias correction term, where f_k are the final sampling weights of the reference survey adjusted for the nonresponse in U_B. We assume that the y_k and δ_k are observed for $k \in s$. Breidt and Opsomer (2017) denote the (4.5) as a difference estimator and consider the estimator (4.5) based on statistical non-parametric learning techniques such as Kernel methods and regression-tree (Hastie, Tibshirani and Friedman, 2001). In the latter case, the estimation process follows these steps: *i*) the survey-weighted regression tree method is applied to the second survey data $\{(y_k, \mathbf{a}_k): k \in s \subset U\}$ where $\mathbf{a}_k$ represents the auxiliary variable value vector observed in the Big Data source; *ii*) a partition of covariate space in *H strata*, denoted as Endogenous Post Strata (Breidt and Opsomer, 2008), is defined as

$$\tilde{\mathbf{a}}_k = \left[1_{\{\tau_{h-1} < \xi(z_k) \leq \tau_h\}} \right]_{h=1}^{H}$$

where the $\{\tau_h\}_{h=0}^{H}$ are known breakpoints; *iii*) $\tilde{y}_k$=$\tilde{\mathbf{a}}_k'\widehat{\boldsymbol{B}}$ is computed, where $\widehat{\boldsymbol{B}}' = \left(\frac{N_1}{\widehat{N}_1}, \ldots, \frac{N_h}{\widehat{N}_h}, \ldots, \frac{N_H}{\widehat{N}_H}\right)$ with $\widehat{N}_h = \sum_{k\in h}(1/\pi_k)$. Breidt and Opsomer (2017) introduce in the discussion the use of *random forests* (Breiman, 2001) instead of a tree-based method without a definitive conclusion. Tipton, Opsomer and Moisen (2013) show empirical evaluations of the (4.5) when using the random forest.

5 Variance estimators

DI estimators are design based. For variance estimation, standard linearisation methods (Särndal et all., 1992) or replication methods (Wolter, 2007) for the regression estimator can be applied.

Pseudo-calibration estimator is model-based. We propose to use a replication method. Specifically, we can apply the Delete a Group Jackknife (DAGJK) method (Kott, 2001; Kott, 2006) which is suitable for treating very large sample.

The DAGJK defines G random replication groups drawn from the parent sample, i.e U_B. Then, G estimation processes are carried out using the sample data without the units of one random replication group.

For the difference estimator (4.5) we apply two independent DAGJK variance estimations respectively for the two components of the estimator. Since U_B and s_B are independent samples the variance of the difference estimator is equal to the sum of the variances of its two components.

The estimation process does not consider the re-computation of the $\tilde{y}_k$.

6 Empirical evaluation on European Community survey on ICT usage and e-commerce in enterprises

We implement the above classes of estimator on the real data of the 2018 *European Community Survey on ICT usage and e-commerce in enterprises* (ICT survey) and Internet data scraped from the enterprise websites. The ICT survey's principal aim is to supply users with indicators on Internet connections and usage (website, social media, cloud computing). The target population of the ICT survey is referred to the enterprises with 10 and more persons employed working in the industry and non-financial market services. The frame population is the Italian Business register (Asia) updated to 2 years before the survey reference period. For the 2018 ICT survey, this population has 199,914 units. The design is a stratified sampling. Four classes of the number of persons employed (0-9; 10-19; 20-249; 250 or more), economic activities (24 Nace groups) and geographical breakdown (21 administrative regions at NUTS 2 level) define the strata. The strata including the fourth size class (the enterprises with 250 and more persons employed) are take-all. The number of units in these strata are 3,342. The 2018 sample of respondents is of

22,097 units. The 2018 ICT survey asked the enterprise, among others, if a) *the website gives the possibility to make online ordering or reservation or booking*; *b*) *there are links to social media on the website*. We refer to the two questions as WEBORD and WEBSM variables. The current ICT survey estimator is a calibration estimator. It calibrates on the number of enterprises and persons employed by economic activity, size class and administrative region according to a complex combination of these variables. We uses the Internet data as Big Data sources. We start with the text documents collected by a web-scraping procedure from the enterprises websites. In particular, we have 93,848 ($= N_B$) scraped websites. Note that the total number of websites in target population is unknown. The ICT survey estimate is 134,655.82 enterprises with a relative error of about 1% (Table 6.1). The web-scraping step returns information retrieval for the WEBSM variable. That means we observe the variable with $y_k = 1$ when the website has a link to a social media and with $y_k = 0$ otherwise. Using the text document of each website we predict with a machine learning technique (Random Forest) the WEBORD variable Bianchi *et all.,* 2018; Bianchi and Bruni, 2019), . That means we predict the variable with $\tilde{y}_k = 1$ when the website has online ordering or reservation or booking functionalities and with $\tilde{y}_k = 0$ otherwise. The prediction is a value in the interval [0; 1]. Righi *et all.* (2019) give insights on the ICT survey and web-scraping and machine learning procedure.

6.1 Estimators

We compare a simplified version of the ICT survey estimator, denoted as T0, with three different RegDI estimators (T1, T2, T3) and three model-based pseudo calibration estimators (M1, M2, M3). T0 calibrates on the number of enterprises and employed persons for four enterprise size classes (0-9; 10-19; 20-249; +249) and for three macro-regions (aggregation of NUTS 2 regions, Centre, North and South). We have $\mathbf{x}_k = (1, e_k)'$ being e_k the number of employed persons in the unit k. The T1 calibration variables are $(\mathbf{x}_k'\boldsymbol{\lambda}_k'; \delta_k\boldsymbol{\lambda}_k')$ and it calibrates on $\mathbf{X}_{(d)} = \sum_U \mathbf{x}_k\, \lambda_{k(d)}$ and $N_{B(d)} = \sum_U \delta_k\, \lambda_{k(d)}$. The T2 calibration variables are $(\mathbf{x}_k'\boldsymbol{\lambda}_k'; \delta_k\boldsymbol{\lambda}_k'; \delta_k\tilde{y}_k\boldsymbol{\lambda}_k')$ and it calibrates on $\mathbf{X}_{(d)}$, $N_{B(d)}$ and $\sum_{U_B} \tilde{y}_k\, \lambda_{k(d)}$. The T3 calibration variables are $(\mathbf{x}_k'\boldsymbol{\lambda}_k'; \delta_k\boldsymbol{\lambda}_k'; \delta_k y_k\boldsymbol{\lambda}_k')$ and it calibrates on $\mathbf{X}_{(d)}$, $N_{B(d)}$ and $\sum_{U_B} y_k\lambda_{k(d)}$. The T4 calibration variables are $(\mathbf{x}_k'\boldsymbol{\lambda}_k'; \delta_k\boldsymbol{\lambda}_k'; \delta_k\tilde{y}_k\boldsymbol{\lambda}_k',\ \delta_k y_k\boldsymbol{\lambda}_k')$ and it calibrates on $\mathbf{X}_{(d)}$, $N_{B(d)}$, $\sum_{U_B} \tilde{y}_k\lambda_{k(d)}$ and $\sum_{U_B} y_k\lambda_{k(d)}$. The M1 estimator calibrates the weights on the estimated totals of enterprise and number of employed persons for four size classes and three macro-regions. We use the estimates of T0 of the above totals. The M1 corresponds to the estimator (4.2) for WEBSM and to the estimator (4.4) for WEBORD. The M2 and M3 are difference estimators for WEBORD total. The f_k in M2 is the sampling calibrated weight adjusted by the factor $\sum_r z_k \,/\, \sum_r \delta_k$, with $z_k = 1$ when the enterprise has the website and $z_k = 0$ otherwise. The M3 estimator uses the factor $\sum_r z_k w_k^s \,/\, \sum_r \delta_k w_k^s$ where w_k^s is the calibrated sampling weight of the ICT survey estimator.

6.2 Results

The estimates at the national level (Table 6.1) gives us some preliminary results. The T1 estimator has not effect on the Coefficient of Variation (CV) of the estimates with respect to the T0. The T2 and T3 estimators reduce the CV for the variable involved in the calibration. We have to consider the T4 estimator for improving the standard errors of both WEBORD and WEBSM variables. The M1 estimator gives two main results: *i*) the two estimates are outside the 95% Confidence Interval (CI) of T0. We have to understand if this is a bias evidence or not; *ii*) the CIs of both estimates are very narrow. We apply the difference estimators, M2 and M3, for the WEBORD total estimate. The value is inside the T0 estimator CI. We can assume to have adjusted the bias for the measurement error of the Big Data target variable. Still, the CV increases with respect to M1 but it is smaller than the CV of T0 and the other DI estimators. As far as the bias of WEBSM total is concerned, Table 6.1 shows that M1 is consistent with T3 and T4 estimators that are design-unbiased. We could assume that T0 produces a downward WEBSM estimation.

Table 6.1: Estimates at the national level

Esti-mator	*Variable*	*Total*	*CI(95%) Lower bound*	*CI(95%) Upper bound*	*Estimate not in T0 CI(95%)*	*CV*
T0	WEB	134,655.82	131,831.46	137,480.18		1.07%
	WEBORD	26,451.41	24,473.67	28,429.14		3.81%
	WEBSM	68,221.35	65,157.69	71,285.01		2.29%
M1	WEBORD	30,120.58	29,956.38	30,284.78	**	0.27%
	WEBSM	79,123.88	78,625.52	79,622.24	**	0.31%
M2	WEBORD	26,860.18	25,740.40	28,361.63		2.47%
M3	WEBORD	26,817.45	26,009.59	27,625.31		1.54%
T1	WEBORD	27,150.30	25,092.20	29,208.40		3.87%
	WEBSM	70,520.33	67,388.36	73,652.30		2.27%
T2	WEBORD	27,387.05	25,806.85	28,967.25		2.94%
	WEBSM	70,684.85	67,577.39	73,792.32		2.24%
T3	WEBORD	28,313.23	26,225.65	30,400.82		3.76%
	WEBSM	77,021.37	74,646.39	79,396.34	**	1.57%
T4	WEBORD	27,541.93	25,989.47	29,094.39		2.88%
	WEBSM	77,022.19	74,647.43	79,396.96	**	1.57%

We compare the estimates by size class domains (Figure 6.1) and macro-regions domains (Figure 6.2). The DI estimator CIs always overlap the T0 estimator CI. The length of CIs looks similar even though the DI CIs are a little bit smaller for some domains (size class 0-9 for WEBORD and WEBSM). The pseudo-calibration estimators gives the shortest intervals. For some domains, the WEBSM estimates are significantly different from the T0 (0-9 size class, Center and North macro-regions). The difference estimator adjusts the WEBORD estimates that are within the T0 estimator CI or at least the CIs of the two estimators overlap. Figures 6.1 and 6.2 include the Tb estimator which is a naïve pseudo-calibration estimator defined as $\left(\widehat{N}_W/N_B\right)\sum_{U_B} y_k^*$, where $\widehat{N}_W$ is the survey-based estimate of the number of units with the website. Table 6.3 and 6.4 investigates the sampling errors of the estimators of the cross-classified domains size class by macro-region (12 domains). We

consider two groups of domains: six domains with a sample size between 344 and 547 units (Group 1) and six domains with a sample size between 1,558 and 8,299 sample units (Group 2). Table 6.2 and Table 6.3 show the average CV (%) respectively for WEBORD and WEBSM. The findings point out that the pseudo-calibration estimator are more efficient.

Figure 6.1: Estimator CIs (95%) by size class for WEBORD total (right) and WEBSM total (left).

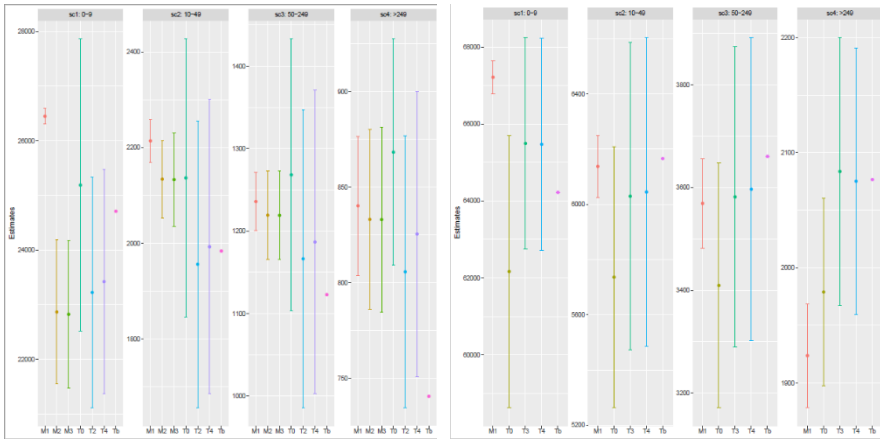

Figure 6.2: Estimator CIs (95%) by macro-regions for WEBORD total (right) and WEBSM total (left).

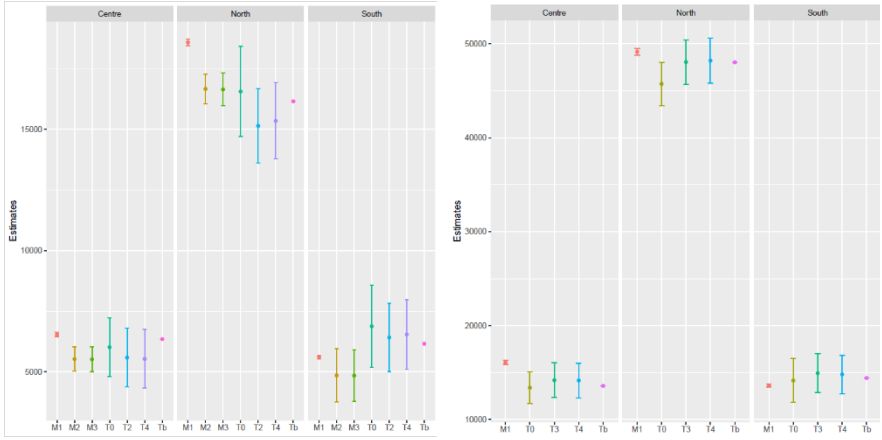

Table 6.2: CV of the estimators for size classes by macro region domain of WEBORD total

Domains	***Average CV(%)***						
	T0	**M2**	**M3**	**T1**	**T2**	**T3**	**T4**
Group 1	12.91	6.18	6.11	13.59	14.36	13.95	14.54
Group 2	7.50	3.73	3.75	7.85	6.26	7.68	6.23

The DI estimators are more efficient than T0 for Group 2 (large domains). Instead, the average of the CV for Group 1 (small domains) is greater than T0.

We explain these findings with the increased number of calibration constraints some units to end up with extreme weights, which will lead to the production of higher variance estimates. This effect is more evident in the small sample size domains.

Table 6.3: CV of the estimators for size classes by macro region domain of WEBSM total

Domains	***Average CV(%)***					
	T0	**M1**	**T1**	**T2**	**T3**	**T4**
Group 1	8.00	3.18	8,47	8.58	9.16	9.26
Group1 2	4.78	1.05	7,02	4.75	3.59	3.59

7 Conclusions

Big Data sources properly used can improve the accuracy of the estimates. In this paper, we introduce, discuss and compare two classes of estimators exploiting the information coming from a Big Data source. The first class takes the Big Data as a source of auxiliary variables into account while a probabilistic survey sample collect the target variables. When the Big Data variables are strictly correlated with survey target variables, the design-based estimates can benefit and the standard errors have a large reduction. The inference approach of these estimators, referred to as Data Integration, is model-assisted. Estimation bias is in the background and depends on the nonresponse issues affecting the survey.

The second class of estimators changes the role of the Big Data. In this case, we directly use the Big Data variables for producing the estimates. Big Data source is a non-probabilistic sample and a probabilistic survey sample focused on the same target population (reference survey) supports the inference. The reference survey needs to: deal with the selection bias of the non-probabilistic survey; adjust the estimates when we have a measurement error on the Big Data target variables. The inference approach of these estimators, referred to this paper as pseudo-calibration estimators, is model-based. Nonetheless, the estimators of this class apply a calibration procedure and the model diagnostic is quite reduced. Variance estimation is computed by means of a replication method. The pseudo-calibration estimators can be biased due to a model failure. On the other hand, the pseudo-calibration estimators increases the real sample size, because they use the non-probabilistic Big Data sample size and the sampling errors can be much smaller than the sampling error of reference survey. The pseudo-calibration estimator sampling errors increase with measurement errors in the Big Data target variables. Both the class of estimators rely on the calibration procedure fostering the practical applicability in the NSIs, in which an automatic estimation process like calibration facilitate the production of the statistics. The experimentation on survey data shows that the sample size make the difference on the sampling errors. The pseudo-calibration estimators based on a large non-probabilistic sample have the best results in terms of precision even though we have to evaluate carefully the risk of bias.

References

1. AAPOR (2015). Big Data in Survey Research. AAPOR Task Force Report. Public Opinion Quarterly, 79, 839–880.
2. Breidt. F.J., Opsomer. J D. (2008). Endogenous poststratification in surveys: Classifying with a sample-fitted model. Annals of Statistics, 36, 403–427.
3. Breidt. F.J., Opsomer. J.D. (2017). Model-Assisted Survey Estimation with Modern Prediction Techniques. Statistical Science, 32, 190–205.
4. Breiman L. (2001). Random Forests. Machine Learning, 45, 5-32.
5. Bianchi G., Bruni R., Scalfati F. (2018). Identifying e-Commerce in Enterprises by means of Text Mining and Classification algorithms. Mathematical Problems in Engineering, Vol. 2018, n. 7231920.
6. Bianchi G., Bruni R. (2019). Website Categorization: a Formal Approach and Robustness Analysis in the case of E-commerce Detection. Expert Systems with Applications.
7. Buelens B., Daas P., Burger J., Puts M., van den Brakel J. (2014). Selectivity of Big data. Discussion Paper nr. 11. Statistics Netherlands.
8. Elliott M., Valliant. R. (2017). Inference for nonprobability samples. Statistical Science, 32, 249–264.
9. Chang C.-C., Lin C.-J. (2001). Training v-support vector classifiers: Theory and algorithms. Neural Computation, 13(9), 2119-2147.
10. Chen Y., Li P., Wu C. (2020). Doubly Robust Inference With Nonprobability Survey Samples. Journal of the American Statistical Association, 2011-2021,
11. Dever. J., Valliant. R. (2010). A comparison of variance estimators for post-stratification to estimated control totals. Survey Methodology, 36, 45–56.
12. Dever. J., Valliant. R. (2016). GREG estimation with undercoverage and estimated controls. Journal of Survey Statistics and Methodology, 4, 289–318.
13. Deville, J. C., Särndal, C. E., (1992). Calibration Estimators in Survey Sampling. Journal of the American Statistical Association, 87, 367-382.
14. EUROSTAT (2018). Report describing the quality aspects of Big Data for Official Statistics. Work Package 8 Quality Deliverable 8.2, ESSnet Big Data.
15. Hastie T., Tibshirani R., Friedman J. (2001). The Elements of Statistical Learning: Data Mining. Inference and Prediction. Springer. New York.
16. Kim J.K. (2022). A gentle introduction to data integration in survey sampling. The Survey Statistician, 19–29.
17. Kim, J. K. and Tam, S. (2021). Data integration by combining big data and survey sample data for finite population inference. International Statistical Review, 382–401.
18. Kim J.K., Rao J.N.K. (2012). Combining data from two independent surveys: a model-assisted approach. Biometrika, 85–100.
19. Kim J.K., Wang Z. (2019). Sampling techniques for big data analysis in finite population inference. International Statistical Review, 177-191.
20. Kott, P. (2006). Delete-a-group variance estimation for the general regression estimator under poisson sampling, Journal Official Statistics, 759–767.
21. Kott, P. (2001). Delete-a-group jackknife. Journal Official Statistics, 521–526.
22. Little. R.J.A., Rubin. D.B. (2002). Statistical Analysis with Missing Data. New York: Wiley.
23. Meng X-L. (2018). Statistical Paradises and Paradoxes in Big Data (I): Law of Large Populations, Big Data Paradox, and the 2016 US Presidential Election. The Annals of Applied Statistics, 12, 685–726.
24. Righi P., Bianchi G., Nurra A., Rinaldi M. (2019). Integration Of Survey Data And Big Data For Finite Population Inference In Official Statistics: Statistical Challenges and Practical Applications. Statistica & Applicazioni, 135-158
25. Särndal C.-E., Swensson B., Wretman J. (1992). Model Assisted Survey Sampling. Springer. New York.
26. Tipton J., Opsomer J., Moisen G. (2013). Properties of endogenous post-stratified estimation using remote sensing data. Remote Sensing of Environment, 139, 130–137.
27. UNECE (2014). A Suggested Framework for the Quality of Big Data. Deliverables of the UNECE Big Data Quality Task Team, December 2014.
28. Wolter, K. (2007) Introduction to Variance Estimation. Springer, London.