A new estimator for integrating the ICT survey data and the information collected in the enterprises websites.

Abstract: The Big Data expands the range of sources that have the potential to be used for Official Statistics and represents an effective reply to the declining response rates and the rising costs of conducting surveys, offering, in the meanwhile, potentially more timeliness and granular statistics. The use of these non-survey data sources generates a paradigm shift: from designed data to data-oriented or data-driven statistics. Therefore, it is necessary to determine under which conditions the these sources makes valid inference on the finite target population. Several statistical and quality frameworks on Big Data have this objective. Nevertheless, they are defined according a general perspective. The paper aims to concretize these frameworks going into detail about the statistical tools to apply in each phase of the data generating process. Our proposed approach relies on combining information from multiple data sources with standard or innovative procedures and makes an integrated and coordinate use of the methods. A real example of use of Big Data in Official Statistics shows how to create the conditions to define a process for obtaining accurate and consistent estimates.

Keywords: Big Data acquisition, Selectivity, combining data sources, machine learning prediction, projection estimators.

1. Introduction

Recently National Statistical Institutes (NSIs) have been using several types of data sources in the production process of Official Statistics, including designed data sources such as censuses and sample surveys, and found data sources such as administrative and transactional data. 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 increasing capability of these technologies to provide digital trails. These sources commonly referred 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 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 purpose of this paper is to elaborate on the current statistical frameworks encompassing Big Data, to stress the criticalities affecting the accuracy of the final statistics and to offer the approaches to

preserve quality standards. The proposal of the paper suggests the reshaping of the known statistical tools leveraged in innovative contexts, to combine information from the multiple imperfect data sources (surveys, administrative and Big Data sources) to model the Big Data selection bias, the survey non-response mechanism and the distribution of the key variables of interest and certain marginal distributions.

The paper covers all the phases of the process. It highlights, in several occasions, the need to have a statistical framework that integrates all its elements: a) the data sources have to represent different components of a unique informative system; b) the data mining techniques for processing the Big Data (for instance, natural language processing or image manipulation procedures, machine learning techniques, etc.) have to be planned coherently; c) data analytic methods implemented in different phases (i.e. machine learning techniques and estimators) have to define a comprehensive toolbox; d) the heterogeneity of the target parameters demands different estimators that must define a system of consistent statistics.

A concrete example of use of Big Data in Official Statistics clarifies these issues. Since the 2017, the Italian Statistical Institute provides experimental statistics using Internet data for reproducing some estimates currently computed by the European Community Survey on ICT usage and e-commerce in enterprises (ICT survey). The paper describes the estimation process step-by-step, it stresses the key phases of the process and how to deal with them to reduce the risk of negative effects. Section 2 introduces the general statistical framework that we consider in the paper and section 3 completes the framework definition describing a class of the target parameters commonly estimated in the Official Statistics. Section 4 defines the estimators. Section 5 is devoted to the application of the statistical framework. The focus is on the complete estimation process from the data acquisition to the estimate. The estimates leveraging Internet data are compared with the survey estimates to evaluate the bias. Finally, section 6 gives a conclusive discussion.

2. General description of the Data Generating-Process and quality issues related to the accuracy

According to AAPOR (2015) proposal the statistics produced by Big Data sources can be described by Data Generating-Process (DGP) based on three phase: Generate, Extract Transform and Load (ETL) and Analyze (Figure 1.1).

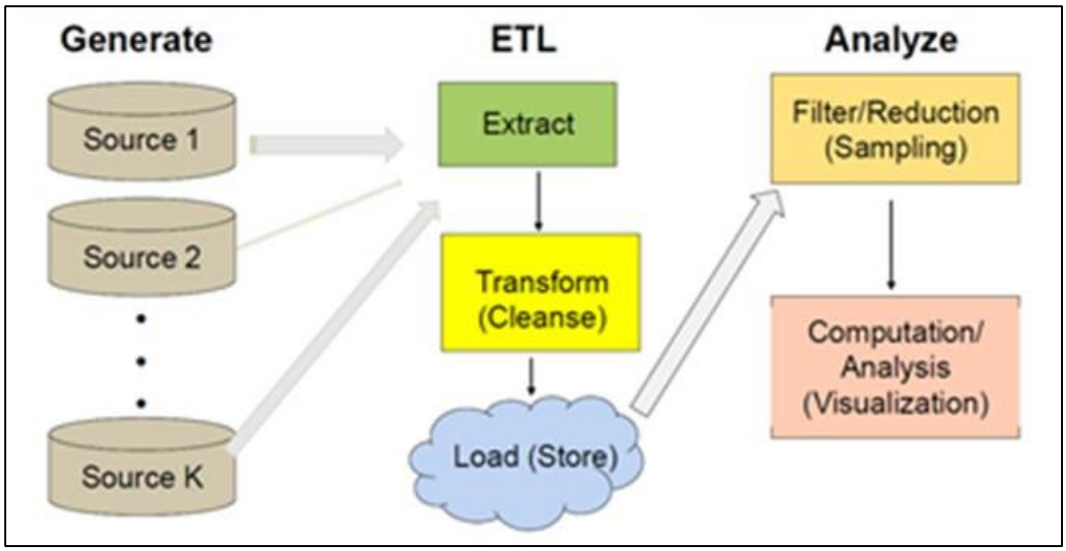

Figure 1.1. Data Generating-Process with the use of Big Data sources*
*AAPOR (2015)

The DGP framework is quite general and for better understanding each phase, we exploit a concrete example. In 2017, Istat started to implement an automatic collection of information from the web to enhance the estimates of website related variables for an enterprise target population. The purpose is to detect services offered by the enterprises just observing their websites without carrying out the survey sampling.

In this context, the Generate phase of the DGP is based on these main steps:

- website address acquisition (acquisition of Uniform Resource Locator -URL):
  - accessing to the administrative sources listing the URL;
  - performing batch queries on the search engines by means of the enterprises identification characteristics available in the statistical business register (URL retrieval with machine learning techniques).
- enterprises identification: to verify if the identified website address belongs to the enterprises of interest.

The output of the Generate phase ends with a structured or unstructured data set, depending on the type of Big Data source. The process enters in the ETL phase in which the following techniques are employed:

- web-scraping techniques for web data acquisition: text scraping, screenshot acquisition, OCR and logo detection;
- natural language processing techniques: finding the meaning of the free text in order to obtain structured data using several techniques such as:

- tokenization: cutting string into still useful linguistic units using string splitting (whitespaces) or regular expressions;
- lemmatization: given a word, its inflectional ending is removed in order to return the word to its basic lemma. This allows to group together the different inflected forms of a word (e.g. plurals of nouns, tenses of verbs, etc.) so they can be analysed as a single item;
- Part-Of-Speech recognition (POS tagging): every word is identified as a particular part of speech (such as noun, verb, etc.).

The ETL phase text analysis involves several text processing tasks:

- language identification;
- information retrieval: collecting the variables of interest directly observables;
- information extraction: extracting structured information from unstructured and/or semi-structured machine-readable documents by using a machine learning approach;

Data extracted by using these tasks are stored in a terms-document matrix that will be used in the next phase.

The analyze phase is arranged with the following steps:

- collecting the variables of interest directly observables in the website (in case of information retrieval);
- predicting the variables of interest not directly observables by means of a predictive model, generally with a non-parametric Machine Learning (ML) techniques (in case of information extraction);
- making inference: producing statistics to the enterprise target population from the sub-population of enterprises with web scraped websites.

Some statistical outputs of the process can be: point estimates of single or composite variables (totals and means), joint distribution estimates (contingency tables).

NSIs carefully supervise the output quality in accordance with the standard quality framework. The paper focus on the accuracy of the statistical output, which is one of the quality dimensions. Accuracy is usually characterized and decomposed into bias (systematic error) and variance (random error) components, having low accuracy in presence of large bias and/or variance. Accuracy may also be described in terms of the major sources of error that potentially can undermine it. This is the aim of the Total Survey Error approach (TSE; Groves and Lyberg, 2010). The errors involved in the TSE are generally common in traditional surveys as well as in surveys using administrative data and Big Data. Nevertheless, for the latter, others types of errors may occur since the process chain should be

longer and more complex than traditional ones. The AAPOR introduced the Big Data Total Error (BDTE) framework (AAPOR, 2015) as an extension of the TSE and states "[TSE] … is quite limited because it makes no attempt to describe the error in the processes that generated the data. In some cases, these processes constitute a "black box" and the best approach is to attempt to evaluate the quality of the end product". AAPOR identifies the errors in each phase of the data generating-process. The paper is interested on the errors affecting the accuracy of statistical output. To give concrete examples of these errors we continue to use the parallel with the information extracted from the enterprises websites by web-scraping process in order to produce statistics related to services offered by the enterprises.

In the Generate phase, we can have:

- erroneous enterprise identification;
- missing data for technology reasons: i.e. the website technology does not allow the scraping process;
- missing data for voluntary reasons; the website architecture blocks the automatic scraping process;
- selective source: the website address acquisition fails, we not able to reach the entire set of enterprises with the website.

Analyze phase can include the following errors:

- ML prediction errors of the interest variable not directly collect from the website;
- prediction errors of the final estimator to make the sample (the set of enterprises with scraped website) representative of the target population;
- Sampling errors related to the partial observation of the target population.

In particular, the missing data weakens the representativeness of the Big Data sample (following the example the set of scraped enterprises' websites) 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."

On the other hand, erroneous values in the Analyze phase introduce bias whether the inference approach is model-based (Valliant *et al*., 2000) or variance in the model-assisted approach (Särndal *et al*. 1992). Finally sampling errors introduces the variance of the estimates model-assisted approach.

3. Parameters of interest

Let U be the target population of size N and let $\boldsymbol{y}$ be the random variable of interest being y_k, the value of the variable for kth unit (k=1, … , N). Let $U_B \subset U$ be the sub-population of size N_B in which the values of $\boldsymbol{y}$ are collected or predicted using the Big Data source. Furthermore, let $U_{\bar{B}}$ be the set of units without information from Big Data source being $U_B \cup U_{\bar{B}} = U$ and $U_B \cap U_{\bar{B}} = \emptyset$. Finally, let $\mathbf{z}_k = (z_k^1, \dots, z_k^p, \dots, z_k^P)'$ be the value vector of the P variables of interest on the kth unit such that $\mathbf{z}_k$ is observed only in a survey referred to the same target population. We assume the y and z have zero/one values. The target parameters are:

$$\bar{Y} = \frac{1}{N}\sum_U y_k \qquad (mean), \tag{3.1}$$

$$\bar{C}_{ij} = \frac{1}{N}\sum_U y_k(i) z_k^p(j) \qquad (cell\ ij\ of\ a\ 2 \times 2\ contingency\ table)\,, \tag{3.2}$$

being

$$y_k(i) = \begin{cases} 1, if\ y_k = i; \\ 0, otherwise, \end{cases}$$

where $i \in \{0; 1\}$ and

$$z_k^p(j) = \begin{cases} 1, if\ z_k^p = j; \\ 0, otherwise, \end{cases}$$

where $j \in \{0; 1\}$,

$$\bar{W} = \frac{1}{N}\sum_{U_1} (y_k \times z_k^*) \qquad (mean\ of\ composite\ variable), \tag{3.3}$$

$$z_k^* = \begin{cases} 1, if\ z_k^1 \times \dots \times z_k^p \times \dots z_k^P > 0; \\ 0, otherwise. \end{cases}$$

4. Estimation process

The parameter (3.1) is the relative frequency of a variable that we can (partially) collect or predict by the use of the Big Data source, while the parameters (3.2) and (3.3) make use of survey data too, so we distinguish the estimation process for the two classes of parameters. Nevertheless, the estimates must be consistent each other in a single estimation system.

4.1 Estimation procedure for $\bar{Y}$

The estimation procedure based on the Big Data information has to take into account two issues:

A) the representativeness of U_B with respect to U;

B) the use of ML predicted values, denoted as $\tilde{y}_k$, instead of y_k in U_B.

A) The representativeness of U_B with respect to U: let δ_k be the U_B membership indicator variable with $\delta_k = 1$ if $k \in U_B$ and $\delta_k = 0$ otherwise, and assume in this phase that y_k is directly observed when $\delta_k = 1$. The assumption will be relaxed at point *ii*). We are interested in estimating $\bar{Y}$ using the Big Data. We can compute the estimator

$$\bar{Y}_B = \frac{1}{N_B}\sum_U \delta_k y_k$$

with error given by

$$\bar{Y}_B - \bar{Y} = \frac{N}{N_B} Cov(\delta, y),$$

being $Cov(\delta, y) = \frac{1}{N}\sum_U\left[(\delta_k - \bar{\delta})(y_k - \bar{Y})\right]$, where $\bar{\delta} = \frac{1}{N}\sum_U \delta_k$. Meng (2018) denotes $\rho_{\delta y} = [\frac{Cov(\delta,y)}{(\sigma_\delta \sigma_y}])$, with ($\sigma_\delta = \sqrt{\frac{1}{N}\sum_U\left[(\delta_k - \bar{\delta})^2\right]}$ and $\sigma_y = \sqrt{\frac{1}{N}\sum_U[(y_k - \bar{Y})^2]}$, the *Data Defect Correlation* (DDC) and with $E_\delta\left[\rho_{\delta y}^2\right]$ the *Data Defect Index* (DDI). Under the Bernoulli model generating the sample U_B with inclusion probability $pr(\delta_k = 1) = \frac{N_B}{N}$, for $k \in U$ the model expectation $E_\delta(\bar{Y}_B - \bar{Y}) = 0$, since the $E_\delta[Cov(\delta, y)] = 0$ (assuming y as not random variable under the model). In a more general (uncontrolled) selection model we can have $pr(\delta_k = 1|y_k = 1) \neq pr(\delta_k = 1|y_k = 0)$ so that DDI is not null, and $E_\delta(\bar{Y}_B - \bar{Y}) \neq 0$ (indicating selection bias). In the paper, we assume the Bernoulli model fails and we focus on a more general selection model. To deal with the selection bias of U_B and making the resulting analysis valid, a first approach assumes that the selection mechanism of the Big Data sample is akin to MAR mechanism (Little and Rubin, 2007) i.e.,

$$pr(\delta_k = 1|y_k = 1, \mathbf{x}_k, \boldsymbol{\lambda}) = pr(\delta_k = 1|y_k = 0, \mathbf{x}_k, \boldsymbol{\lambda}) = pr(\delta_k = 1|\mathbf{x}_k, \boldsymbol{\lambda}) \; \forall \; k \in U, \qquad (4.1)$$

where $\mathbf{x}_k$ is a vector of values of the auxiliary variables known for each $k \in U$, $\boldsymbol{\lambda}$ is unknown parameter. Denoting with $w_k = 1/pr(\delta_k = 1|\mathbf{x}_k, \boldsymbol{\lambda})$, an unbiased estimator is given by (Meng, 2018)

$$\hat{\bar{Y}}_B = \frac{1}{N_B}\sum_U \delta_k y_k w_k.$$

A method to estimate the $w's$ is to use a *reference survey*¸ *s*, in parallel to the Big Data sample (Elliott and Valliant, 2017). The reference survey selects a random sample from the same population and the vector $(\delta_k, y_k, \mathbf{x}_k)$ is observed for $k \in s$. In this general setting Kim and Wang (2019) propose to estimate the $pr(\delta_k = 1|\mathbf{x}_k, \boldsymbol{\lambda})$ using a pseudo-likelihood approach based on a parametric generalized linear model estimated with the reference survey data. The estimator is denoted *propensity score weighting estimator*.

Elliott and Valliant (2017) propose the *quasi-randomization inferential approach* in which the basic block is the computation of the *pseudo-weights, w's*. The computation requires the reference survey and applies the Bayes rule. Both methods aim to compute the inclusion probability such that the selection mechanism for Big Data is non-informative (DDI=0).

In this paper we propose another approach. We do not consider U_B as a sample of U. Instead, we define a statistical framework where U is a take all sample (census) with $pr(\delta_k = 1) = 1$ for $k \in U$. The sample U is affected by a kind of unit non-response and the inclusion probabilities of the respondents, the units in U_B, are adjusted for reducing nonresponse bias. We assume the nonresponse follows a MAR mechanism conditionally to the $\mathbf{x}_k$ vector, and apply a calibration step based on the following optimization problem:

$$\begin{cases} min \sum_{U_B} d(p_k, w_k) \\ \sum_{U_B} \mathbf{x}_k \, w_k = \mathbf{X} \end{cases} . \qquad (4.2)$$

where $d(.)$ is a convex function, denoted as distance function, (Singh and Mohl, 1996), p_k=1 is the initial weight, w_k is the unknown weight, $\mathbf{X} = \sum_U \mathbf{x}_k$ is a vector of totals, that we assume as known or estimated by large and accurate survey (e.g., Dever and Valliant, 2010, 2016) with $\mathbf{x}_k$ known for each $k \in U_B$. The reference surveys can be the source for estimating the totals. The solution gives a unique set of w weights and the estimator is given by

$$\hat{\bar{Y}}_{PC,B} = \frac{1}{N} \sum_U \delta_k y_k w_k \qquad (4.3)$$

being $w_k = 0$ when $\delta_k = 0$. The optimization problem in (4.2) is solved applying the calibration algorithm (Deville and Särndal, 1992). We denote $\hat{\bar{Y}}_{PC,B}$ as a *pseudo-calibration estimator* and underline that the purpose of the process is to correct the selection bias.

Remark 4.1. The proposed estimator has simple and straight implementation. It leverages well known and widely used statistical tools in the NSIs. It combines information from the multiple data sources (administrative data, survey and Big Data sources).

Remark 4.2. If we assume that the condition (4.1) is valid and, furthermore, if we have $f(\boldsymbol{y}_{U_B}|\mathbf{x}_k, \boldsymbol{\theta}) = f(\boldsymbol{y}_{U_{\bar{B}}}|\mathbf{x}_k, \boldsymbol{\theta})$, being $f(.)$ the density function, with $\boldsymbol{y}_{U_B}$ and $\boldsymbol{y}_{U_{\bar{B}}}$ respectively the random variables in U_B and $U_{\bar{B}}$ and $\boldsymbol{\theta}$ a parameter vector, we can predict the value of y_k for $k \in U_{\bar{B}}$, ignoring the selection mechanism, and compute the model-unbiased estimator

$$\hat{\bar{Y}}_B^* = \frac{1}{N}\left[\sum_{U_B} y_k + \sum_{U_{\bar{B}}} \hat{y}_k\right] \qquad (4.4)$$

being $\hat{y}_k$ the predicted value for $k \in U_{\bar{B}}$. In case the conditional density is parameterized with a linear model the estimator (4.4) and (4.3) coincide and the $\hat{\bar{Y}}_{PC,B}$ and it is a design- based, model based unbiased estimator.

Remark 4.3. The proposed estimator can consider the *propensity score weights* by Kim and Wang or the *pseudo-weights* by Elliot and Valliant as initial weights in the optimization problem. In this case the purpose of the process is to enhance the precision of the estimates, while initial weights guarantees the unbiasedness.

Remark 4.4. The p_k values do not affect the optimization problem solution. However, with particular $d(.)$ function, the optimization problem is solved by means of an iterative process and setting $p_k = N/N_B$ should guarantee a faster convergence to the optimal solution.

Remark 4.5. The existence of the optimal solution depends on the $d(.)$ function. The calibrated weights could contain smaller than 1 or negative values. There are $d(.)$ functions constraining the calibrated weights to be in a prefixed range of values but in these cases the convergence could be not achieved. For the discussion about convergence and the admissible solutions of the optimization problem (4.2) see, for example, Deville and Särndal (1992).

B) The use of $\tilde{y}_k$ instead of y_k in U_B. The web-scraping can substantially collect the y_k variable (information retrieval) or can achieve structured information (information extraction) for implementing a ML technique to computed the prediction, $\tilde{y}_k$, of the variable of interest on the *k*th unit (k=1, … , N_B). In the latter case, the estimator (4.3) has to be refined plugging-in the $\tilde{y}_k$ synthetic values for y_k ,

$$\hat{\bar{Y}}^{P}_{PC,B} = \frac{1}{N}\sum_{U} \delta_k \tilde{y}_k w_k, \qquad (4.5)$$

where $\tilde{y}_k$ is null for $\delta_k = 0$. The estimator (4.5) assumes the form of the *projection estimator*. Kim and Rao (2012) define a model assisted framework of the estimator (4.5) 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 define the conditions in order to have unbiased estimates. When the conditions are not satisfied, an unbiased estimator is

$$\hat{\bar{Y}}^{D}_{PC,B} = \hat{\bar{Y}}^{P}_{PC,B} + \frac{1}{N}\sum_{s \subset U} \frac{y_k - \tilde{y}_k}{\pi_k}, \qquad (4.6)$$

in which the second term of the right hand side of the (4.6) is the bias correction term, where π_k is the inclusion probability of the reference survey. We denote the (4.6) as *difference estimator*.

Breidt and Opsomer (2017) consider the estimator (4.6) 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

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

where the $\{\tau_h\}_{h=0}^{H}$ are known break points; *iii*) $\tilde{y}_k$=$\tilde{\mathbf{a}}_k'\widehat{\boldsymbol{B}}$ is computed, where $\widehat{\boldsymbol{B}}' = \left(\frac{N_1}{\widehat{N}_1}, \dots, \frac{N_h}{\widehat{N}_h}, \dots, \frac{N_H}{\widehat{N}_H}\right)$ with $\widehat{N}_h = \sum_{k\in h}(1/\pi_k)$. Breidt and Opsomer (2017) introduce in the discussion the use of the *random forests* (Breiman, 2001) instead of tree-based method without a definitive conclusion. Tipton, Opsomer and Moisen (2013) show empirical evaluations of the (4.6) when using the random forest. Montanari and Ranalli (2005) propose the use of neural network techniques.

4.2 Estimation procedure for $\bar{C}_{ij}$ and $\overline{W}$

The $\bar{C}_{ij}$ and $\overline{W}$ are functions of survey variables not observed in the Big Data. The estimation procedure can follows two approaches: *i*) the mass imputation (Park and Kim, 2019, Yang and Kim, 2018); *ii*) the model-assisted estimator (Deville and Särndal, 1992).

The former approach foresees to impute the value of the $\mathbf{z}_k$ vector for all $k \in (U \cap \bar{s})$. The choice of this approach, regardless the statistical properties, requires a deep change in the production process that can be less acceptable in the NSIs. We describe the second approach that is more familiar and acceptable for the official statistic agencies. The model-assisted approach uses a new calibration estimator applied to the survey data in which we add new calibration constraints related to the estimated produced by the Big Data. The initial sample weights, $1/\pi_k$, are adjusted obtaining by the final weights, g_k, such that

$$new\ constraints \rightarrow \begin{cases} \sum_s \mathbf{x}_k\, g_k = \sum_U \mathbf{x}_k \\ \sum_s \mathbf{y}_k\, g_k = N\hat{\bar{Y}} \end{cases} \qquad (4.7)$$

where, we assume that the constraints $\sum_s \mathbf{x}_k\, g_k = \sum_U \mathbf{x}_k$ are currently used for calibrating and adjusting for unit non-response the survey estimator, and $\hat{\bar{Y}}$ is the estimate achieved with one among the estimators (4.3), (4.4), (4.5) or (4.6).

Remark 4.6. The estimator of the parameters $\bar{C}_{ij}$ and $\overline{W}$ requires a complete integration between Big Data and survey data. Both the sources support each other and the estimator produces consistent

statistics with respect to the estimator of $\bar{Y}$. The statistical framework integrates data sources and processes in accordance with the purposes given in section 1 of the paper.

5. Empirical evaluation on *European Community Survey on ICT usage and e-commerce in enterprises*

We implement the general statistical framework on real data of the 2017 *European Community Survey on ICT usage and e-commerce in enterprises* (ICT survey) and Internet data scraped from the enterprise websites. This is an upgrade of the framework proposed in 2017 Istat experimental statistics (Barcaroli et al. 2018) accessible at Istat website: https://www.istat.it/en/archivio/216641 (accessed on October 2019). Here we introduce the unbiased difference estimators, the estimation of the contingency tables and the mean of composite variables.

5.1 The survey data

The ICT survey is part of European Community statistics on the information society and its principal aim is to supply users with indicators on Internet connections, on Internet usage (website, social media, cloud computing), on electronic integration of the business process (i.e. through the use of software to interact and share business information internally like ERP, CRM or externally with other enterprises of value chain), on eCommerce (electronic sales and purchases), eInvoice and on more innovative ICT investments (Robotics, Internet of Things, Artificial Intelligence, Big data analysis). ICT survey is also one of the major yearly sources of data for the Digital Agenda Scoreboard and contributes to composite Digital Economy and Society Index (DESI) used to summarize the progress of the European digital economy.

The target population of ICT survey is referred to the enterprises with 10 and more persons employed working in 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 2017 ICT survey, this population is of 184,865 unit. The sampling design is the following: *i*) a census for the "with 250 and more persons employed" enterprises (3,152 enterprises); *ii*) a stratified simple random sample for the smaller and medium enterprises (10-249 persons employed). The stratification variables are: 4 classes of number of persons employed, economic activities (24 Nace groups) and geographical breakdown (21 administrative regions at NUTS 2 level). The sample size is of 32,361 enterprises with a response rate of 66.2%. The 2017 sample of respondents is of 21.410 legal units.

Member States and Eurostat choose every year the special section of the questionnaire to be investigated more deeply to respond to the political need of digital progress measuring. For this reason, the questionnaire changes every year as well as the tabulation program. The parameters of

interest are simple indicators based on answers given to each qualitative questions (positive and negative indicators on ICT usages) and indicators derived from more than one questions defining composed indicators.

The 2017 ICT survey asked to the enterprise, among others, if a) *the website gives the possibility to make online ordering or reservation or booking* (e_webord); b) *there are job advertisements in the website* (e_webjob); c) *there are links to social media in the website* (e_websm).

In the following we assume for the generic variable, that $y_k = 1$ when the answer of the unit k is positive and with $y_k = 0$ otherwise.

Istat supplies aggregated estimates on simple distributions of these variables for economic activity by size class and administrative region. Furthermore, the survey produces the estimates of the frequency of some composite variables. For example, the e_*webcom* variable defined as follows: $y_k = 1$ if the website has web ordering facilities and at least one of the following website functionalities: *to have the description/price list of goods or services*; *to customize or design the products*; *to personalized content*; *to track the order*. The questionnaire collects these four yes/no variables.

Finally we consider the estimation of a 2×2 contingency table combining the presence e_webord variable by the variable indicating whether the enterprises *has declared to have made sales through web channels* (*website, app, digital platforms*) *in the previous year* (e_awsell) collected in a specific section of the ICT questionnaire.

The current 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 combinations of these variables.

The estimation process uses the Internet data (scraped data from the websites) for estimating the frequencies of e_webord, e_webjob and e_websm (section 4.1) and survey data for composite variable and contingency table (section 4.2). Table 5.1 shows the process for the target parameters and the relative estimators.

Tab. 5.1 Parameters and estimators making use of the Internet data source

Variable	Type of parameter	Observed Big Data values	Estimator
e_webord	Relative frequency	Predictions	(4.5) - (4.6)
e_webjob	Relative frequency	Predictions	(4.5) - (4.6)
e_websm	Relative frequency	Trues	(4.3)
e_webcom	Composite variable relative frequency	Predictions and survey values	Current survey estimator plus a calibration based on the e_webord estimate constraints (4.7)

e_webord by e_awsell	Contingency table	Predictions and survey values	Current survey estimator plus a calibration based on the e_webord estimate constraints (4.7)

Table 5.1 highlights that automatic collection of information from the website is able to directly observe the variable e_websm, while a prediction is performed for the e_webord and e_webjob variables. The following sections describe the implementation of the statistical framework in each phase of the Generate-data process.

5.2 The URL acquisition.

The beginning step of the estimation process is the enterprise web address acquisition. In our empirical evaluation we consider a) an administrative source (Consodata) listing approximately 90,000 enterprises with URLs; b) the downloading information from some proper thematic directory sites; c) the performing batch queries on the search engines by means of the enterprises identification characteristics (name of the enterprises) available in the statistical business register (URL retrieval with machine learning techniques).

The successive step verifies the correspondence between websites and enterprises. In case of available URL from administrative sources, the procedure check if the URLs are valid and exist at start. It proceeds with the syntactic validation of the strings, the check of the recurring errors and the domain extraction. In case of non-existing URL, it performs a web search using search engines in order to find the most similar URL (at most ten for each enterprise). In case the enterprise web address is not available in advance, the procedure performs batch queries on the search engines by means of the available enterprises identification characteristics. In particular, the name of the enterprise is used as a search string, and then a query on a search engine is performed, collecting the first ten links returned as the result of the query.

At this point, the URL is the output of the retrieval procedure and we take the enterprise identification variables from the website (Fiscal Code, VAT Number, Business Name, Address, etc.) through the use of Information Retrieval techniques and compute a score by the comparison of the scraped information with the same information available in Asia through matching techniques and string similarity metrics (Jaro-Winkler, Levenshtein, etc.). For enterprises with more than one likely link, we take the one with the highest score. A supervised machine learning approach implements the matching. We consider as training set a random subsample of the enterprises of the ICT survey for which the correct URL is available and fit the model in order to predict the exact correspondence between website and enterprise for the remaining cases.

5.3 *ETL phase*

Completed the Generate-phase we perform the automatic extraction of statistical information from Internet. In case of the e_websm variable we are able to exactly collect the values (information retrieval). Instead, for the e_webord and e_webjob variables the process is more articulated and produces predicted or classified values (Figure 5.1).

The processes collects texts (web-scraping), identifies relevant terms (text mining) and models the relationships between these terms and the characteristics we are interested to estimate (classification).

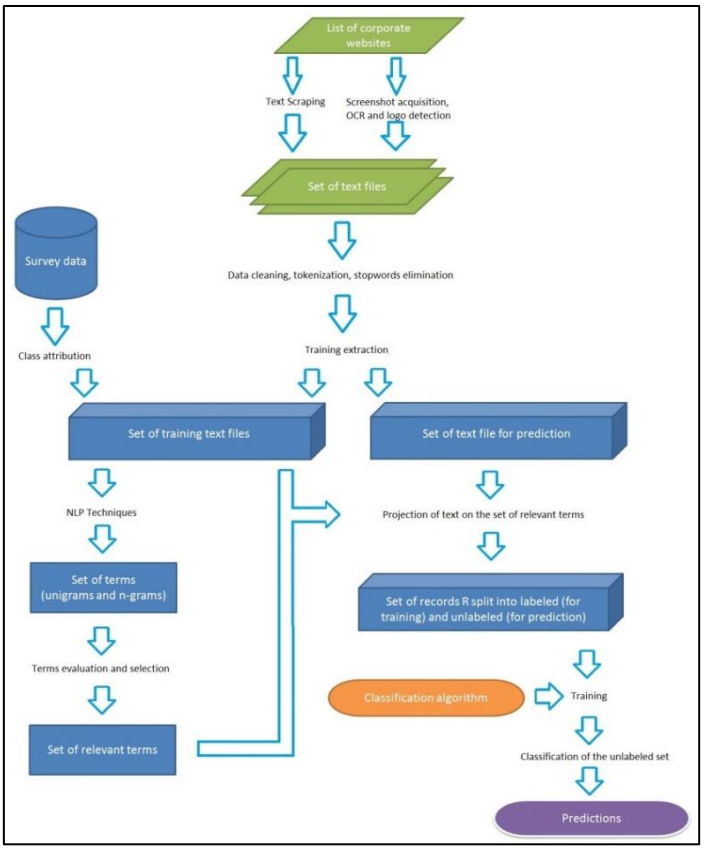

Figure 5.1. The global classification approach

For each enterprise, a web-scraping procedure reads and saves the content of the website. It reads text and some additional information from the homepage to other pages up to a certain depth. The additional information are the attributes of HTML elements, the name of the image files, the keywords of the pages. Moreover, all the types of images that appear in the pages are processed with Optical Character Recognition (OCR) routine, in order to read also the words provided in images and unusual writing techniques (e.g. Flash animations). In fact, these words are particularly relevant (consider for instance the case of logos, `buy' and `pay' commands, etc.). A preliminary step of character

segmentation to find the regions of the image containing text has been carried out using the Marvin open source image processing framework, developed in Java (Archanjo, Andrijauskas and Munoz 2008). The OCR is obtained by using the Tesseract Open Source OCR Engine (Smith, 2007). The output is saved into a NoSQL DBMS and each text website related is arranged in a text file d_k (with $k \in U_B$). The number of words downloaded for each website goes up to 3,000,000. We classify d_k as positive when $y_k = 1$ and negative when $y_k = 0$. The class label have been interactively checked by human intervention with concrete visualization of the website. Since the vast majority of this information is irrelevant to the interest phenomenon in the next step we try to exclude the noise information as much as possible performing an articulated text mining procedure (Bianchi and Bruni, 2018). Initially we clean the text by tokenization and removing all non-alphabetic symbols and the stop-words (articles, prepositions, etc.). Then by using natural language processing techniques we extract the dictionaries: the Ω set of the *uni-grams*, i.e., the single words (performing lemmatization process using the software TreeTagger (Schmid, 1995)) and the Ψ sets of the *n-grams* appearing in the same files (performing part-of-speech recognition process using the software TreeTagger) being the *n-gram* the sequences of *n* adjacent words that are typically used together ("credit card" is a *bi-gram* example). We define up to a maximum value of $n = 5$ *n-grams*. We keep the well-composed *n-grams* (for example, in the case of *bi-grams*, we keep the pairs: noun and verb, noun and adjective, noun and adverb, etc.). A this point we measure the relevance for each term of Ω and Ψ. We exploit a Term Evaluation (TE) function and in particular the Chi Square metrics, denoted as χ^2. We use the χ^2 metrics to measure the dependence between the generic term and the class (positive or negative) of the generic file *d* containing *uni-gram* $\omega \in \Omega$ or the *n-grams* in Ψ (in practice they collapse to $n = 2$ *bi-grams*). The *uni-gram* TE score is given by $\text{score}(\omega) = \chi^2_{\omega+} + \chi^2_{\omega-}$, where $\chi^2_{\omega+}$ is the positive score and $\chi^2_{\omega-}$ is the negative score being

$$\chi^2_{\omega+} = \frac{(n_{\omega+} + n_{\omega} + n_{+} + n)(n_{\omega+}n + n_w n_{+})^2}{(n_{\omega+} + n_w)(n_{+} + n)(n_{\omega+} + n_{+})(n_{\omega} + n)}$$

where $n_{\omega+}$ is the total number of occurrences of distinct word ω in class + files; n_+ is the total number of distinct words in class + files; n_ω is the total number of occurrences of distinct word ω in all files; *n* is the total number of occurrences of all distinct words. The negative score is defined similarly, except that all the above values are computed for negative files. The analogous score function is defined for the *n-grams*.

We take terms that have a TE score larger than a predetermined threshold and up to a maximum of terms. Finally, we obtain data record projecting each text file d_k on the set of terms *T*. In this way,

we reduce d_k to a real vector of size $|T|$. Where each real number refers to the number of occurrences in the file. This number is normalized with respect to a measure that depends on the file size.

In the case study each record has 800 terms obtained from *uni-grams*, 200 terms from *n-grams*, and one class label.

Accomplished the ETL-phase, we proceed to fit the model or classifier (machine learning technique) using the survey weights on the training set for predicting the values of target variables for all the enterprises for which the retrieval and scraping of their websites was successful. Note that the use of the weights responds to the inferencial requirement of the estimator (4.4) or (4.5). In our case, we have performed preliminary tests with several classifiers (see also Bianchi and Bruni, 2015) by means of *Scikit learn- package* (Pedregosa et al., 2011) included into scientific Phyton distributions. The best results have been obtained with Random Forest (Ho, 1998; Breiman, 2001), Support Vector Machines (SVMs; Vapnik, 1995; Chang & Lin, 2001) and Logistic regression model (Agresti, 2002; Freedman, 2009).

On these three techniques we perform the training phase on the training sets performing "grid-search" on classifier parameters using *3-fold cross-validation* obtaining the sets of the predictive models. In particular, we draw a random sample of 4,755 websites from the ICT sample and check interactively the class (positive or negative) of the websites. To perform the classification task, we select a training set of 2,377 websites. Various combinations of parameters values are analysed and the one with the best cross-validation accuracy is picked. Note that the use of a grid search approach is standard in practical applications, and even though it cannot theoretically guarantee to determine exactly the optimal parameters, it is generally regarded as a very reasonable compromise between time and performance. The extraction have been randomly performed 3 times, and all performance results are averaged on the 3 trials. The remaining websites 2,378 are used as test set to accuracy of the classification procedure. Finally, by knowing the real class of the records in the above test sets, we compute the *confusion matrix* and we use its elements (True Positives - TP, False Negatives - FN, True Negatives - TN, False Positives - FP) to evaluate the following performance measures:

- Accuracy *a*, defined as the percentage of correct predictions over all predictions:

$$a = \frac{100(TP + TN)}{TP + FN + TN + FP}$$

- Precision *p*, also called the positive predictive value, defined as the percentage of true positive records in all positive predictions: $p = \frac{100\,TP}{TP+FP}$
- Sensitivity φ, also called the true positive rate, defined as the percentage of correct positive predictions in all real positive records: $\varphi = \frac{100\,TP}{TP+FN}$

- F1-score (Sokolova, Japkowicz and Szpakowicz, 2006), which is the harmonic mean of precision and sensitivity:

$$F1 = \frac{200\,TP}{2\,TP + FP + FN}$$

The F1-score appears to be the most relevant performance measure, since it fully evaluates the correct identification of the positive records, that is the most important and difficult task, and because it has low sensitivity to data imbalance. Table 5.2 shows the main performance indicator connected with the e_webord variable.

Table 5.2. Performance indicator for the three compared learners on the test set for predicting the e_webord variable

Learner	**Accuracy**	**Recall**	**Precision**	**F1-score**
Logistic regression model	0.88	0.64	0.66	0.65
SVMs	0.90	0.62	0.76	0.68
Random Forest	0.90	0.72	0.74	0.73

By analyzing the results, we observe that Random Forest classifier provides the best performances in our experiments. However, we notice that the SVMs and Logistic regression model produce slightly lower results. These good results depend on the accuracy of the previous text-mining phase. An interesting work that analyzes the robustness of each technique with respect to the presence of misclassified training records is described in (Bianchi and Bruni, 2019).

5.3 *Analyze phase: inference*

ETL phase returns information retrieval for the e_websm variable while the random forest predicts the variable e_webord and e_webjob after the information extraction steps. We have two levels of prediction: the probability of $y_k = 1$, or the assignment of the synthetic value 0 or 1. The process proceeds to compute the estimates. For sake of brevity, here we focus on the estimates related to the variable e_webord. The process begins to investigate the selection bias of the U_B sample of the scraped websites. The current survey based estimation of $\bar{Y}$ is equal to 14.97% and the estimator $\bar{Y}_B$ gives a values equal to 21.13% outside the 95% Confidence Interval (CI). We think that $\bar{Y}_B$ estimator is biased. For applying the propensity score weighting estimator or the quasi-randomization inferential approach we verify whether the hypothesis $pr(\delta_k = 1|k \in s) = pr(\delta_k = 1|k \in \bar{s})$ holds. The survey has 21,410 respondents with 16,632 enterprises declaring to have a website and 13,532 scraped websites (81.4%). The official estimated number of enterprises with website in target

population is 133,361 and the Generate-phase succeeds in scraping for 85,464 websites (64.1%). The two percentages underline that the probability to be scraped varies conditionally in being in the ICT survey or not. The reason is the following: the ICT reference survey asks to the enterprise the URL. On the other hand, for the units that are not in the reference survey nor in administrative URL archive, the URL retrieval must be carried out. This operation pushing downward the probability to have a scraped website. Then we apply the pseudo-calibration estimator. The pseudo-calibrated weights of the 85,464 units reproduce the number of enterprises and the number of persons employed by 24 economic activities, 4 macro economic sectors by 4 size classes and 21 administrative regions according the Italian Business Register. We compute four estimators:

a) the projection estimator (4.5) using the estimated probability of $y_k = 1$, hereinafter projection estimator (a) type;
b) the projection estimator (4.5) using the predicted 0 or 1 value for y_k, hereinafter projection estimator (b) type;
c) the difference estimator (4.6) with $H = N_B$ i.e. using directly the estimated probability of $y_k = 1$, hereinafter difference estimator (c) type;
d) the difference estimator (4.2) with H=2 (0 and 1), hereinafter difference estimator (d) type.

The bias correction terms of the estimator (c) and (d) have been calculated adjusting the direct weights by the inverse probability to be scraped conditionally to have the website.

Then we verify whether the respective estimates belong to the CI of the current estimates of the ICT survey for three types of estimation domains: 24 economic activities, 4 economic macro sectors by 4 size classes and 21 administrative regions. Table 5.2 shows the number of domain in which the estimates are outside the CI.

Table 5.2. Number of estimates using internet data outside of the coefficient interval of the current estimates

Domain of estimates	**Number of domains**	**Estimator (a) type**	**Estimator (b) type**	**Estimator (c) type**	**Estimator (d) type**
Macro sector by size classes	16	3	2	2	1
Economic activity	24	7	4	6	3
Administrative Region	21	3	3	3	3
Total	**61**	**13**	**9**	**11**	**7**

The difference between the (a) type and (b) type projection estimators with the respective bias corrected versions (c) type and (d) type seems to improve the data quality. In particular, the estimator (a) type produce 13 estimates outside the CI and this number decreases to 11 for the estimator (c) type; the estimator (b) type has 9 estimates outside the CI and the estimator (d) type has 7 estimates external to the intervals. We note that the out of the CI estimates of the (c) type estimator includes the analogous set of the (a) type estimator and the same relationship holds between the estimator (d) and (c) type. That indicates that the bias correction term moves the estimates in the CI and it never moves the estimates outside the CI. We have similar results for the variable e_webjob. Table 5.3 gives a general estimation result, in particular for the (a) type estimator, for the e_webord, e_webjob, The e_websm relative percentage frequency uses the (4.3) estimator,

Tab. 5.3 Relative percentage frequencies of the e_webord, e_webjob with the current estimator and the type (a) projection estimator (e_websm uses (4.3) estimator)

Variable	**Current estimator**	**Confidence Interval**		**Estimator using Internet data**
		Lower bound	Upper bound	
Web ordering offered functionalities (e_webord)	14.97	13.81	16.13	15,51
Job advertisement in the website (e_webjob)	10.78	10.02	11.53	13.91
Presence of social media links in the website (e_websm)	31.25	29.90	32.60	36.68

The estimation process for the e_webord by e_awesel contingency table utilizes the calibration estimator using the (4.7) constraints, since we cannot perform the information retrieval or extraction for the e_awesel variable. Applying the current estimator (Table 5.4 A) we can observe that the e_webord marginal distribution is inconsistent with the estimated distribution exploiting the internet data (Table 5.3).

Tab. 5.4 Relative percentage frequency contingency table estimated with the current estimator (A part) and with the type (a) projection estimator (B part).

	A Current calibration estimator				B Calibration estimator with new constraints (4.6)		
Variable	e_webord=0	E_webord=1	Total		e_webord=0	E_webord=1	Total

e_awsell=0	83.23	6.87	90.10		82.46	7.49	89.05
e_awsell=1	1.80	8.10	90.90		2.03	8.02	10.05
Total	**85.03**	**14.97**	**100.00**		**84.49**	**15.51**	**100.00**

The calibration estimator including the constraints referred to e_webord total estimates at different domain levels corrects the contingency estimates and the marginal distribution goes towards the estimates with the internet data (Table 5.4 B).

Finally, we investigate the e_webcom distribution estimates. Table 5.5 shows in columns (1) the absolute frequency of e_webord estimates with the projection estimator (a) type and in column (2) the e_webcom absolute frequency with the current survey estimator by economic activity. The difference (1)-(2) should be positive by definition while in some cases do not occurs. Table 5.5 column (4) presents the calibration estimator including the new constraints. The column (5) shows now only positive difference between e_webord and e_webcom absolute frequencies. The proposed estimator produce consistent statistics with the estimates based on the internet data.

6. Conclusion

The NSIs begin to exploit new data sources for producing official and experimental statistics by now and the scientific community has being proposing new statistical and quality frameworks for establishing the conditions to make valid inference. The paper describes the approaches and defines the statistical tools in order to fulfill these conditions. In details, we investigate the selectivity concern of the Big Data and the approaches to deal with it in different informative contexts creating the conditions to have representativeness of the Big Data. We focus on methods proposed in literature and suggest a slightly different approach based on well-known statistical techniques. Furthermore, we face the estimation problem when the Big Data source offers predictive values of the target variable instead of collecting the true values. We consider the entire data generating-process to guarantee the conditions that allow to make valid inference hold and we provide an approach to have a system of consistent estimates based on different estimators. All these points are concrete problems in Official Statistics. The integration concept is always in the background in our framework. Integration by using: a reference survey for correction the selection bias or the machine learning prediction bias; the administrative source for calibrating and returning a representative data. The framework and solutions are supplied in a general form, the empirical evaluation realizes, with practical examples, how to implement the proposed framework.

Finally, the paper covers especially the bias issue when using the Big Data but we underline that the proposed estimators have defined the variance estimators with analytic expression or with replicated method (i.e. bootstrap). Further analysis are planned in the next developments.

Tab. 5.5 Frequency estimates of $y = 1$ for the e_webord (with projection estimator (a) type) and e_webcomp (with current survey estimator and calibration estimator with (4.6) constraints) variable by 24 economic activities

Economic activity	**Projection Estimator (a) type**	**Current estimator**		**Calibration estimator with new constraints**	
	e_webord	**e_webcom**	**Difference**	**e_webcom**	**Difference**
	(1)	**(2)**	**(3)=1)-(2)**	**(4)**	**(5)=(1)-(3)**
Manufacture of food products. beverages and tobacco products	1270.3	1385.9	-115.5	1196.7	73.6
Manufacture of textiles. apparel. leather and related products	1465.1	1433.2	31.9	1240.0	225.1
Manufacture of wood and paper products. and printing	682.8	596.4	86.4	620.7	62.1
Manufacture of coke and refined petroleum products. of chemicals and chemical products. of basic pharmaceutical products and preparations. of rubber. plastic and of other non-metallic mineral products	996.6	735.6	261.0	806.2	190.4
Manufacture of basic metals and fabricated metal products. except machinery and equipment	1036.3	737.4	299.0	947.5	88.9
Manufacture of computer. electronic and optical products	165.0	126.8	38.2	152.3	12.6
Manufacture of electrical equipment and of machinery and equipment n.e.c.	1166.0	608.7	557.3	1113.7	52.3
Manufacture of transport equipment	202.0	268.3	-66.3	193.3	8.7
Manufacture of furniture. other manufacturing. and repair and installation of machinery and equipment	964.9	722.7	242.1	957.0	7.9
Electricity. gas steam. air conditioning supply. water supply. sewerage. waste management andremediation activities (D. E)	322.1	196.9	125.3	175.7	146.4
Construction	1054.3	363.6	690.7	674.6	379.7
Wholesale and retail trade and repair of motor vehicles and motorcycles	7435.7	6784.6	651.1	6631.1	804.6
Transport and storage. except warehousing and support activities for transportation (H except53)	1453.1	1022.4	430.7	797.9	655.2
Postal and courier activities	31.0	57.2	-26.2	30.4	0.6
Accommodation	4235.9	4398.3	-162.5	3807.5	428.4
Food service activities	3081.5	1919.8	1161.8	1934.1	1147.4
Publishing activities	219.0	268.9	-49.9	204.7	14.3
Motion picture. video and television programme production. sound recording	142.0	94.3	47.8	139.7	2.3
Telecommunications	56.0	55.5	0.5	42.5	13.6
IT and other information services	639.0	408.4	230.5	568.3	70.7
Real estate activities	87.0	69.3	17.7	81.1	5.9
Professional. scientific and technical activities except veterinary activities (M except 75)	817.0	346.2	470.8	646.0	171.0
Administrative and support service activities except travel agency. tour operator and other reservation service and related activities (N except 79)	895.7	435.2	460.5	749.5	146.2
Travel agency. tour operator and other reservation service and related activities	251.0	241.5	9.5	231.4	19.6

References

AAPOR (2015). Big Data in Survey Research. AAPOR Task Force Report. *Public Opinion Quarterly. 79*. pp. 839–880.

Agresti A (2002). *Categorical Data Analysis*. Wiley.

Archanjo G.A., Andrijauskas F., Munoz D. (2008). Marvin A Tool for Image Processing Algorithm Development. *Technical Posters Proceedings of XXI Brazilian Symposium of Computer Graphics and Image Processing*.

Bianchi G., Bruni R. (2015). Effective Classification using Binarization and Statistical Analysis. *IEEE Transactions on Knowledge and Data Engineering* 27(9). 2349-2361.

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. 2018.

Bianchi G., Bruni R. (2019). Website Categorization: a Formal Approach and Robustness Analysis in the case of E-commerce Detection. *Expert Systems with Applications.* DOI 10.1016/j.eswa.2019.113001

Breidt. F. J., Opsomer. J. D. (2008). Endogenous poststratification in surveys: Classifying with a sample-fitted model. *Ann*. *Statist*. 36 403–427.

Breidt. F. J., Opsomer. J. D. (2017). Model-Assisted Survey Estimation with Modern Prediction Techniques. *Statistical Science*. 32 . 190–205.

Breiman. L. (2001). Random forests. *Mach*. *Learn*. 45. 5–32.

Breiman L. (2001). Random Forests. *Machine Learning.* 45. 5-32.

Buelens B., Daas P., Burger J., Puts M., van den Brakel J. (2014). Selectivity of Big data. *Discussion Paper nr. 11*. Statistics Netherlands.

Elliott. M., Valliant. R. (2017). Inference for nonprobability samples. *Statistical Science*. 32. 249–264.

Chang C.-C., Lin C.-J. (2001) Training *v*-support vector classifiers: Theory and algorithms. *Neural Computation*. 13(9). 2119-2147.

Dever. J., Valliant. R. (2010). A comparison of variance estimators for post-stratification to estimated control totals. *Survey Methodology.* 36. 45–56.

Dever. J., Valliant. R. (2016). GREG estimation with undercoverage and estimated controls. *Journal of Survey Statistics and Methodology.* 4 289–318.

Deville, J. C., Särndal, C. E., 1992, *Calibration Estimators in Survey Sampling*. *JASA*. 87. 367-382.

EUROSTAT (2018). *Report describing the quality aspects of Big Data for Official Statistics*. Work Package 8 Quality Deliverable 8.2, ESSnet Big Data.

Freedman D.A. (2009). *Statistical Models: Theory and Practice*. Cambridge University Press.

Groves. R. M., Lyberg. L. (2010). Total Survey Error: Past Present and Future. *Public Opinion Quarterly. 74.* 5. pp. 849-879.

Hastie. T., Tibshirani. R., Friedman. J. (2001). The Elements of Statistical Learning: Data Mining. *Inference and Prediction*. Springer. New York.

Ho T.K. (1998) The Random Subspace Method for Constructing Decision Forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 20(8). 832-844.

Kim. J. K., Rao. J. N. K. (2012). Combining data from two independent surveys: a model-assisted approach. *Biometrika*. 99. 85–100.

Kim J.K., Wang Z. (2019). Sampling techniques for big data analysis in finite population inference. *International Statistical Review*. 87. 177-191.

Yang S., Kim J.K. (2018).Integration of survey data and big observational data for finite population inference using mass imputation, *arXiv preprint arXiv:1807.02817*.

Little. R. J. A., Rubin. D. B. (2002). *Statistical Analysis with Missing Data*. New York: Wiley.

Loh W.-Y. (2014) Fifty years of classification and regression trees. *International Statistical Review*. 82. 329-348.

Meng X-L. (2018). Statistical Paradises nd 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.

Montanari. G. E., Ranalli. M. G. (2005). Nonparametric model calibration estimation in survey sampling. *JASA*. 100. 1429–1442.

Park S., Kim J.K. (2019). Mass imputation for two-phase sampling. *Journal of the Korean Statistical Society*. Accepted.

Pedregosa F. et al. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*. 12. 2825-2830.

Särndal. C.-E., Swensson. B., Wretman. J. (1992). *Model Assisted Survey Sampling*. Springer. New York

Schmid H. (1995) Improvements in Part-of-Speech Tagging with an Application to German. *Proceedings of the ACL SIGDAT-Workshop*. Dublin. Ireland.

Singh, A. C., Mohl, C. A., 1996, Understanding Calibration Estimators in Survey Sampling. *Survey Methodology*. 22. 107-115.

Smith R. (2007). An Overview of the Tesseract OCR Engine. in *Proc. of the Ninth International Conference on Document Analysis and Recognition*. 629-633. 2007 ISBN:0-7695-2822-8 IEEE Computer Society. Washington. USA.

Sokolova M., Japkowicz N., Szpakowicz S. (2006). Beyond Accuracy. F-Score and ROC: A Family of Discriminant Measures for Performance Evaluation. In *Advances in Artificial Intelligence. Lecture Notes in Computer Science*: Sattar A., Kang B. (eds) AI 2006. Vol. 4304. Springer. Berlin. Heidelberg

Tipton. J., Opsomer. J., Moisen. G. (2013). Properties of endogenous post-stratified estimation using remote sensing data. *Remote Sens. Environ.*,139. 130–137.

UNECE (2014). *A Suggested Framework for the Quality of Big Data*. Deliverables of the UNECE Big Data Quality Task Team December, 2014.

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

Vapnik V. (1995). *The Nature of Statistical Learning Theory*. Springer.