Joint use of sampling data and Big Data: the experience with the Istat survey on the use of ICT by enterprises


(*) Italian National Institute of Statistics (ISTAT)
(**) CINECA
Outline

- The “Big Data representativeness problem”
- The Istat survey on the use of ICT by enterprises: the design based approach
- Internet data as additional or alternative source
- Use of web scraping, text mining and machine learning techniques to implement a model based approach
- Adjustment of the model based approach to take into account the representativeness problem
- Design and model based estimators: a compared evaluation
One of the most important problems in using Big Data for producing official statistics is their representativeness with respect to the population of interest.

In general, the population $B$ of Big Data covers a subset of the target population $U$ w.r.t. the statistician is interested to measure a variable $y_U$.

There is a first problem of undercoverage of target population by using Big Data (overcoverage is not a problem): we call $R$ the “censoring” process that makes only a part of $y_U$ measurable with Big Data.

On $B$ we can observe values $z_B$.

We can select a sample $S$ in order to observe the $y_S$ values together with the $z_S$.

We call $I$ the sampling process used to select the sample.
If the two processes of censoring and sampling are such that they can be considered as \textit{ ignorable}, then

\begin{equation}
    p(y_U | z_B, I, R) \approx p(y_U | z_B)
\end{equation}

In other terms:

\textbf{if} the subset of the target population for which Big Data are available is not significantly different from the whole target population, 

\textbf{and}

\textbf{if} the sampling is such that the sample is representative of the whole target population,

\textbf{then}

the scientific process of translating the Big Data observations \( z_B \) into the measurement of interest \( y_U \) can be considered valid for the whole target population.
The Istat Survey on ICT in Enterprises

To illustrate a scenario based on the combined use of survey data and Big data, we take into consideration a given survey, the «Survey on the use of ICT by Enterprises», carried out in all Member States of the European Union.

In Italy, the survey, based on a stratified sampling design, investigates on a universe of about 182,000 enterprises with at least 10 employees, by means of a sample involving 32,000 of them, of which 61% are respondents.
The Istat Survey on ICT in Enterprises

This is a subsection of the questionnaire:

<table>
<thead>
<tr>
<th>Use of a Website or Home Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>B7. In January 2013, did your enterprise have a Website or Home Page? (Filter question)</td>
</tr>
<tr>
<td>B8. In January 2013, did the Website or Home Page have any of the following?</td>
</tr>
<tr>
<td>a) Online ordering or reservation or booking, e.g. shopping cart</td>
</tr>
<tr>
<td>b) A privacy policy statement, a privacy seal or certification related to website safety</td>
</tr>
<tr>
<td>c) Product catalogues or price lists</td>
</tr>
<tr>
<td>d) Order tracking available online</td>
</tr>
<tr>
<td>e) Possibility for visitors to customise or design the products</td>
</tr>
<tr>
<td>f) Personalised content in the website for regular/repeated visitors</td>
</tr>
<tr>
<td>g) Advertisement of open job positions or online job application</td>
</tr>
</tbody>
</table>

Optional

In the 2016 round of the survey, more than 14,000 declared a website and indicated related URL.
Web scraping + text processing + machine learning

Reference population: 182,000 enterprises

Population frame (ASIA)

Sample selection

32,000 enterprises

Web scraping

10,000 websites

Texts

14,000 URLs

Data collection on 19,000 enterprises

Microdata

Text mining and machine learning

Predictors

Big Data: Internet as Data Source

Websites and social networks

e-commerce
e-recruitment
e-tendering
…
Web scraping and text processing

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>ragso</td>
<td>ANNO</td>
<td>Sito_esteso</td>
</tr>
<tr>
<td>14397563</td>
<td>2015</td>
<td><a href="http://www.farmaciachappano.it">www.farmaciachappano.it</a></td>
</tr>
<tr>
<td>10915770</td>
<td>2015</td>
<td><a href="http://www.lottomatica.it/servizi">www.lottomatica.it/servizi</a></td>
</tr>
<tr>
<td>5098916</td>
<td>2015</td>
<td><a href="http://www.energiasrdi.it">www.energiasrdi.it</a></td>
</tr>
<tr>
<td>48858</td>
<td>2015</td>
<td><a href="http://www.magonrispa.it">www.magonrispa.it</a></td>
</tr>
<tr>
<td>2504436</td>
<td>2015</td>
<td><a href="http://www.2m-elettronica.it">www.2m-elettronica.it</a></td>
</tr>
<tr>
<td>19184537</td>
<td>2015</td>
<td><a href="http://www.accademiaunipmn.it">www.accademiaunipmn.it</a></td>
</tr>
<tr>
<td>3973062</td>
<td>2015</td>
<td><a href="http://www.acialetti.it">www.acialetti.it</a></td>
</tr>
<tr>
<td>675255</td>
<td>2015</td>
<td><a href="http://www.accomhotels.com">www.accomhotels.com</a></td>
</tr>
<tr>
<td>19047080</td>
<td>2015</td>
<td><a href="http://www.acsescafatt.it">www.acsescafatt.it</a></td>
</tr>
<tr>
<td>27285481</td>
<td>2015</td>
<td><a href="http://www.adit.it">www.adit.it</a></td>
</tr>
<tr>
<td>18734</td>
<td>2015</td>
<td><a href="http://www.agrimac.it">www.agrimac.it</a></td>
</tr>
<tr>
<td>203204</td>
<td>2015</td>
<td><a href="http://www.agristorei.com">www.agristorei.com</a></td>
</tr>
<tr>
<td>221650</td>
<td>2015</td>
<td><a href="http://www.arhitar-e-werk.com">www.arhitar-e-werk.com</a></td>
</tr>
<tr>
<td>214677</td>
<td>2015</td>
<td><a href="http://www.albergomalta.com">www.albergomalta.com</a></td>
</tr>
<tr>
<td>26445443</td>
<td>2015</td>
<td><a href="http://www.almacontact.it">www.almacontact.it</a></td>
</tr>
<tr>
<td>20920828</td>
<td>2015</td>
<td><a href="http://www.armonia-energy.it">www.armonia-energy.it</a></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>site_descri...</th>
<th>tagged_content</th>
<th>freq_terms</th>
<th>num_terms</th>
<th>content_size</th>
</tr>
</thead>
<tbody>
<tr>
<td>3098915</td>
<td>elettrotecnico componente seguente radio clienti ...</td>
<td>7 9 2 1 6 14 11 3 2 7 2 7 7 4 4 2 2 1 5 7 6 1 6 7 3 ...</td>
<td>53</td>
<td>427</td>
</tr>
<tr>
<td>2504436</td>
<td>nuovo utensile numero personale opera emulsion ...</td>
<td>2 6 6 2 2 5 3 2 5 2 2 1 5 1 6 4 3 5 3 5 5 2 2 2 ...</td>
<td>72</td>
<td>570</td>
</tr>
<tr>
<td>48858</td>
<td>nuovo friuliano padre grande uscire situare compi ...</td>
<td>1 7 2 2 2 3 3 2 4 4 1 2 1 3 7 3 2 3 2 2 4 2 2 6 ...</td>
<td>72</td>
<td>530</td>
</tr>
<tr>
<td>214677</td>
<td>elettrocnico cimiteriale saldatore saldatrice saldau ...</td>
<td>3 2 2 2 2 6 2 1 1 2 2 2 2 1 1 1 2 2 3 4 2 2 3 ...</td>
<td>155</td>
<td>1304</td>
</tr>
<tr>
<td>18464363</td>
<td>elevato occupare cura retribuzione ottimo gestione ...</td>
<td>3 2 8 2 2 4 2 2 1 0 4 1 1 2 7 3 1 0 5 1 0 2 2 0 1 ...</td>
<td>213</td>
<td>1741</td>
</tr>
<tr>
<td>26445443</td>
<td>pelle cura consentire scegliere menu latta latte su ...</td>
<td>6 2 2 1 9 4 6 3 3 4 3 4 3 1 8 2 2 2 2 5 3 6 5 0 4 ...</td>
<td>306</td>
<td>2423</td>
</tr>
<tr>
<td>14397563</td>
<td>fama fax hotel bita design selezionare energia ro ...</td>
<td>7 2 2 2 2 4 3 2 3 2 2 4 4 6 1 6 1 6 2 3 4 5 4 2 1 ...</td>
<td>123</td>
<td>748</td>
</tr>
<tr>
<td>3697298</td>
<td>ufficiale coreografo laboratorio speciale incluso ar ...</td>
<td>5 8 2 3 3 5 2 3 1 1 1 1 1 3 4 2 1 3 1 3 2 1 2 ...</td>
<td>397</td>
<td>3094</td>
</tr>
<tr>
<td>20920828</td>
<td>asta contributo modulo regolamento affidare riser ...</td>
<td>1 4 2 4 2 2 1 2 2 6 2 2 2 2 8 2 4 4 7 2 2 1 5 4 8 4 1 ...</td>
<td>82</td>
<td>650</td>
</tr>
</tbody>
</table>
Many different supervised learning algorithms have been applied:

- Logistic Regression
- Naïve Bayes
- *Ensembles* (Bagging, Boosting, Random Forests)
- Neural Networks
- Support Vector Machines
- Statistical and Logical Analysis of Data
## Performance of machine learning algorithms

<table>
<thead>
<tr>
<th>Learner</th>
<th>Accuracy</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Est. diff.</th>
<th>F1 measure</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Logistic</td>
<td>0.83</td>
<td>0.53</td>
<td>0.89</td>
<td>0.01</td>
<td>0.53</td>
<td>0.01625</td>
</tr>
<tr>
<td>2. Naïve Bayes</td>
<td>0.80</td>
<td>0.46</td>
<td>0.87</td>
<td>0.00</td>
<td>0.46</td>
<td>0.99490</td>
</tr>
<tr>
<td>3. Random Forest</td>
<td>0.83</td>
<td>0.53</td>
<td>0.90</td>
<td>0.01</td>
<td>0.55</td>
<td>0.00006</td>
</tr>
<tr>
<td>4. Bagging</td>
<td>0.82</td>
<td>0.44</td>
<td>0.90</td>
<td>0.03</td>
<td>0.48</td>
<td>0.11520</td>
</tr>
<tr>
<td>5. Boosting</td>
<td>0.81</td>
<td>0.50</td>
<td>0.88</td>
<td>0.00</td>
<td>0.50</td>
<td>0.56530</td>
</tr>
<tr>
<td>6. Neural Net</td>
<td>0.82</td>
<td>0.52</td>
<td>0.89</td>
<td>0.01</td>
<td>0.52</td>
<td>0.10180</td>
</tr>
<tr>
<td>7. SVM</td>
<td>0.83</td>
<td>0.64</td>
<td>0.88</td>
<td>0.01</td>
<td>0.59</td>
<td>0.00018</td>
</tr>
<tr>
<td>8. SLAD</td>
<td>0.84</td>
<td>0.62</td>
<td>0.90</td>
<td>0.01</td>
<td>0.60</td>
<td>0.00018</td>
</tr>
</tbody>
</table>
Current estimation approach (design based)

One of the target estimate of the survey is the «Number of enterprises that have a website and use it for online ordering (web ordering)» considering, as population of interest U, the enterprises having a website.

Under the current approach, this number is estimated by using the calibration estimator:

$$\hat{Y}_{GREG} = \left( \sum_{k \in U} x_k \right) \cdot \hat{\beta} + \sum_{k \in s} d_k \left( y_k - x_k \cdot \hat{\beta} \right) = \hat{Y}_{HT} + (X - \hat{X}_{HT}) \cdot \hat{\beta} = \sum_{k \in s} w_k y_k$$

The estimate is obtained by modifying the initial weights (that depend only on the inclusion probabilities of sampling units) using as auxiliary variables the number of firms and the number of employees, according to the information contained in the Business Register ASIA.

This allows to reduce the bias induced by the high non response rate, as final weights let each respondent unit represent also non respondent units having similar characteristics.
Combined use of survey and Internet data (model based)

Under this approach, there is a combined use of survey data together with the data directly collected on the Internet:

1. All websites used by enterprises included in the target population are subject to scraping, in order to collect the html text contained in them.
2. Text is processed by using text mining techniques, in order to produce a «document-term matrix».
3. Survey data act as the training set, used to tune a machine learning algorithm where the target variable is «e-commerce (yes/no)» and the explanatory variables are the terms in the documents-terms matrix.
4. The algorithm is applied to the subset of population, in order to predict the value of «e-commerce (yes/no)» for all of them.
5. The estimate of «e-commerce (yes/no)» can be then obtained by counting the values «yes» in the whole population of observed and scraped units:

\[
\hat{Y}_{\text{Alg}} = \sum_{k \in s} y_k + \sum_{k \in (U' - s)} \hat{y}_k
\]
Joint use of sampling data and Big Data: the experience with the Istat survey on the use of ICT by enterprises

ITACOSM 2017 – The 5th ITAlian Conference on Survey Methodology
(i) Adjustment for total non response in sample data

Models are fitted on the sample data. Are these data representative of the whole population of interest? Total non response is the result of a non ignorable process.

Models should be fitted taking into account also variables that can explain total non response, by:

- considering the same calibrated weights used in the current estimation procedure (in parametric models i.e. the logistic model);
- considering also those variables as explanatory variables (in non parametric models).
(ii) Adjustment for undercoverage of Big Data population

Actually, we are able to reach only a subset of the population of enterprises having a website: $U' \subset U$ (*undercoverage of target population by using Big Data*).

We can use calibration in order to let Big Data population represent the whole population of interest.

Therefore, the estimates are obtained by calibrating, using the same auxiliary variables used to calibrate sample observations (number of firms and the number of employees, according to the information contained in the Business Register ASIA).
A comparison of estimates

For the 2016 round of the survey, a subset of related estimates have been compared to those obtained by the alternative model-based approach.

This subset includes the proportion of enterprises offering e-commerce in their websites, cross-classified by economic activity (NACE 1 digit, 4 values) and size (4 classes of number of employees).
Joint use of sampling data and Big Data: the experience with the Istat survey on the use of ICT by enterprises
ITACOSM 2017 – The 5th ITAlian Conference on Survey Methodology
Joint use of sampling data and Big Data: the experience with the Istat survey on the use of ICT by enterprises
ITACOSM 2017 – The 5th ITAlian Conference on Survey Methodology
Simulation study

A simulation study has been conducted, on a very realistic basis: all the conditions related to the sampling survey and on the availability of Internet data have been re-created. In particular, varying prediction performance of models have been considered, from the current one to an optimal one.

A comparison between the quality of estimates currently obtained by the statistical survey, and the ones obtainable by the use of Internet data has been carried out, based on the calculation of:

1. bias
2. variance
3. Mean Squared Error (bias$^2$+variance)
Simulation study

Joint use of sampling data and Big Data: the experience with the Istat survey on the use of ICT by enterprises
ITACOSM 2017 – The 5th ITAlian Conference on Survey Methodology
Simulation study

Joint use of sampling data and Big Data: the experience with the Istat survey on the use of ICT by enterprises

ITACOSM 2017 – The 5th ITalian Conference on Survey Methodology
Simulation study

Joint use of sampling data and Big Data: the experience with the Istat survey on the use of ICT by enterprises
ITACOSM 2017 – The 5th ITAlian Conference on Survey Methodology
Simulation study

The simulation has shown that:

1. Comparing adjusted and non adjusted model based estimates, in terms of bias and MSE the former are always better than the latter

2. Comparing model based and design based estimates
   a. in terms of bias and variance, model based are often better than design based
   b. in terms of MSE, the adjusted model based estimates are better than design based with few exceptions.

These conclusions are true only with an acceptable prediction performance of models (F1-measure).
Conclusions

- Big Data sources are, by definition, out of the control of statisticians.
- The most favorable situations are the ones in which other sources (i.e. statistical, administrative) can be used as a “bridge” between Big Data and sources under control.
- Model based estimators are the most suitable in these situations.
- Adjustment procedures in order to take into account the “representativeness problem” should be applied.
- A careful evaluation of quality of estimates obtained by using Big Data sources should always be carried out.
References


Thank you for your attention

barcarol@istat.it
m.scarno@istat.it
natalia.golini@gmail.com
nurra@istat.it
parighi@istat.it
piersimo@istat.it
sesalamo@istat.it