Data integration challenges – ICT survey and Web scraped data from enterprise web sites

Giorgio Alleva

President, ISTAT, Italy

Abstract

A multi-source approach (based on the combined use of survey, administrative and Big Data sources) should allow to overcome the usual limits of each single source. This multi-source approach requires a shift in the paradigm of statistical inference. The traditional one followed by National Statistical Offices is usually based on the design-based survey sampling theory and model-assisted inference. The new one (algorithmic-based inference) is derived by data science: the emphasis is on the exploration of all available data, seeking information that has not been extracted so far.

Istat has experimented this new approach in order to replicate a subset of the estimates currently produced by the sampling survey on "Survey on ICT usage and e-Commerce in Enterprises", yearly carried out by Istat and by the other member states in the EU. Target estimates of this survey include the characteristics of websites used by enterprises to present their business (for instance, if the website offers web ordering facilities; job vacancies; presence in social networks). To produce these estimates, data are collected by means of the traditional questionnaires.

An alternative way is to make use of Internet data, i.e. to collect data by accessing directly the websites, processing the collected information to individuate relevant terms, and modelling the relationships between these terms and the characteristics we are interested to estimate. To do that, the sample of surveyed data plays the role of a training set for fitting models that can be applied to the generality of enterprises owning a website. Administrative data (mainly contained in the Business Register) are used to cope with representativeness problems. The sequential application of web scraping, text mining and machine learning techniques allows to obtain the information suitable for applying a prediction approach and produce estimates that can be compared to the survey ones.

In terms of quality (accuracy), the impact of the new estimators is potentially both positive (reduction of the variability of the estimates, and of the bias due to sampling variance, to total non-response and to measurement errors in the survey) and negative (model bias and variance). Whenever the quality level of estimates obtained by means of this new approach is deemed to be not lower than the ones produced by the traditional process, the former has to be preferred, as it allows not

only to produce aggregate estimates, but also to predict individual values, useful for instance to enrich the information contained in registers.

The evaluation carried out so far (also by means of simulations studies) demonstrate that (i) the alternative set of estimates are compatible with the survey design based ones, as most of the former lay in the confidence interval ranges of the latter and (ii) the variability of estimates is in general lower for model base estimates, while bias is comparable if we take into account the response errors in the survey.

In conclusion, this successful experience has shown that one of the most important Big Data sources, the Internet data, can be harnessed to produce information both at unit level, in order to enrich the Business Register, and at aggregate level, in order to produce estimates whose accuracy and timeliness are competitive with the costly traditional survey based estimates.