Use of Big Data in official statistics

Giulio Barcaroli
Italian National Institute of Statistics (Istat), Italy – barcarol@istat.it

Abstract

More abundant and higher-quality data are required to meet increasingly complex knowledge demands from users and stakeholders in modern societies. At the same time, budget pressures and costs associated with response burden limit the scope for new surveys. Harnessing new data sources in official statistics is therefore decisive to provide an adequate supply of information and knowledge while enhancing the effectiveness of statistical production processes. This paper presents the approach followed, and the results produced, in some experiments carried out by Istat in order to investigate the possibility to produce official statistical information by making use also of the new data sources represented by Big Data. Pilots have been planned by making reference to different possible scenarios: (i) leaving unaltered the classical sampling survey framework, data collection can be based on the availability of Big Data (use of *web scraping* and *text mining* techniques in "ICT in enterprises" survey); (ii) Big Data can be harnessed jointly with survey data, and considered as auxiliary information that can increase the accuracy and timeliness of estimates (*Google Trends* indicators for producing nowcasting and small area estimates); (iii) Big Data are used as the only source in alternative to the one represented by traditional statistical surveys (use of *mobile phone data* to estimate origin/destination matrix of daily mobility for work and study at the spatial granularity of municipalities). Though further research and development are needed, first results are encouraging, and new investigations are being launched: use of data from social networks (Twitter and Facebook) to estimate Consumer Confidence Index, estimation of road traffic flows by using sensors (webcams), use of scanner data for the calculation of Consumer Price Index.

Keywords: web scraping; text mining; Internet queries; mobile phone data; multiple data sources.