An overview of methods for treating selectivity in big data sources

Maciej Beręsewicz[1,2]

[1]Department of Statistics, Poznań University of Economics and Business, e-mail: maciej.beresewicz@ue.poznan.pl
[2]Centre for Small Area Estimation, Statistical Office in Poznań, e-mail: m.beresewicz@stat.gov.pl

Abstract

Official statistics is now considering seriously big data as a significant data source for producing statistics. It holds the potential for providing faster, cheaper, more detailed and completely new types of statistics. However, the use of big data brings also several challenges. One of them is the non-probabilistic character of most sources of big data, as very often they were not designed to produce statistics. The resulting selectivity bias is therefore a major concern when using big data.

This paper presents a statistical approach to big data, searching for a definition meaningful from the statistical point of view and identifying its main statistical characteristics. It then argues that big data sources share many characteristics with Internet opt-in panel surveys and proposes this as a reference to address selectivity and coverage problems in big data. Coverage and the self-selection process are briefly discussed in mobile network data, Twitter, Google Trends and Wikipedia page views data. An overview of methods which can be used to address selectivity and eliminate, or mitigate, bias is then presented, covering both methods applied at individual level, i.e. at the level of the statistical unit, and at domain level, i.e. at the level of the produced statistics. Finally, the applicability of the methods to the several big data sources is briefly discussed and a framework for adjusting selectivity in big data is proposed.

The lecture will be based on the document *An overview of methods for treating selectivity in big data sources* (Beręsewicz *et al.*, 2018). In addition, examples based on real datasets will be provided.

Acknowledgements: This report is the result of a study to identify methods for treating selectivity in big data sources contracted by the European Commission (Eurostat) to SOGETI. The study was conducted by Maciej Beręsewicz and Risto Lehtonen with terms of reference set up by Martin Karlberg, Loredana Di Consiglio and Fernando Reis. The report also benefited from the comments of the members of Eurostat task force on big data and of three external reviewers.

Keywords: Big data, Selectivity

References

Beręsewicz, M., Lehtonen, R., Reis, F., Di Consiglio, L. & Karlberg, M. (2018). *An overview of methods for treating selectivity in big data sources.*