A MODEL-BASED APPROACH TO ESTIMATE BIAS IN INTERNET DATA SOURCES

Maciej Beręsewicz
Poznań University of Economics, Poland, maciej.beresewicz@ue.poznan.pl

New data sources, such as the Internet data sources (IDS) or big data, are gaining recognition not only in the non-statistical but also statistical literature. Several papers and presentations devoted to these new data sources address the problem of representativeness, bias, data quality or, generally, the usefulness of these data for official statistics. Information from those sources is not only used for producing statistics, but is also increasingly being used in the model-based approach as a source of auxiliary variables. This is also true of small area estimation, particularly in the field of area-level models, for example supplied with data from Google Trends (Porter et al., 2014; Rao, 2014). Therefore, it is crucial to evaluate and quantify the bias that can be observed in these data sources.

In view of the above, the main aim of the paper is to present an attempt to indirectly estimate bias of selected characteristics (the price of m^2, number of rooms and floor area) of flats offered to sale on the secondary real estate market in Poland using information from IDS. For this purpose, the author will apply the model-based approach to estimate the bias of statistics based on administrative registers proposed by Zhang (2012). In addition, the paper will present an organization of statistical concepts related to Internet data sources and demonstrate the importance of assessing bias, which tends to be neglected in the non-statistical literature.

Furthermore, an extension of the concept of modelling bias described by Zhang (2012) will be discussed. The proposed approach will adapt a more general linear mixed model that takes into account (i) several data sources, (ii) domain and time series data and (iii) the context of new data sources, in particular IDS. The methods presented in the paper will be exemplified using actual data from real estate portals in Poland. All calculations will be conducted in the **R** statistical package (R Core Team, 2014).

The paper is the part of the research is supported by the National Science Centre, Poland, Preludium 7 grant no. 2014/13/N/HS4/02999.

References

R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/

Porter, A. T., Holan, S. H., Wikle, C. K., & Cressie, N. (2014). Spatial Fay–Herriot models for small area estimation with functional covariates. Spatial Statistics, 10, 27-42.

Rao J.N.K. (2014), Inferential Issues in Model-based SAE: Some New Developments, International conference on SAE 2014, Poznań.

Zhang L.-C. (2012), On the accuracy of register-based census employment statistics, Presentation on European Conference on Quality in Official Statistics in Greece, 2012.