ASSESSING SELECTIVITY AND REPRESENTATIVENESS OF INTERNET DATA SOURCES FOR THE REAL ESTATE MARKET IN POLAND

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

Estimation conducted by National Statistical Institutes heavily relies on statistical data, such as surveys, census or establishments reporting. Recently, in an effort to meet information needs, official statisticians have been searching for new (non-statistical) data sources, including both government (administrative records) and non-government (private) data. The main examples of the latter group are Internet data sources (IDS) or big data, which have recently been widely discussed in the context of official statistics (Buelens et al. 2014; Daas et al., 2015).

The main goal of this paper is to assess selectivity and representativeness of IDS (web-scraped data) for the real estate market (REM) in Poland. The objective will be achieved in three steps: (1) a description of web-scraping as a method of collecting statistical data and an account of co-operation between the University and private companies that held these data; (2) a definition of weak and strong selectivity in the case of aggregated data available; (3) measurement of representativeness defined by comparing trends estimated from the IDS and official statistics. The choice of the real estate market for illustration was motivated by (i) the importance of the REM for the economy, (ii) insufficient information coverage of the REA in Poland and (iii) the importance of the Internet as a source of information on the REA. The study was conducted on the basis of domain (city level) quarterly data for the period between 2012 and 2014 from 5 IDS about the secondary REM. Research conducted by the National Bank of Poland and the Central Statistical Office in Poznań, as well as register data on transactions were used as references. All calculations were made in **R** statistical package (R Core Team, 2015).

Several web-scraping algorithms were developed to enable the continuous monitoring of the real estate market in Poland. These algorithms automatically collected data concerning 16 biggest cities in Poland (capitals of Voivodships, NUTS2 level) directly from selected web-portals. In addition, thanks to co-operation established earlier with three companies owning real estate market web-portals, it was possible to obtain historical and collect current data for the research.

Since access to individual-level data was not possible for all periods, the analysis of *selection bias* at the domain level was conducted by means of a linear mixed model (Zhang, 2012). Based on estimated bias, the author will propose a definition of weak and strong selectivity. Weak selectivity is defined to be present when the data source effect is observed, whereas strong selectivity occurs when there is interaction between domain and a data source. Several cases will be discussed in detail.

Finally, in order to assess representativeness for key variables, such as price for m^2, number of rooms and floor area of flats, a trend approach is proposed. Due to possible bias of IDS-based point estimates, a direct comparison of such levels can be misleading. Therefore, an alternative approach is put forward. It is based on the comparison of estimated trends in IDS and official data to assess whether the changes in time are coherent with official statistics.

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

Buelens, B., Daas, P., Burger, J., Puts, M., van den Brakel, J. (2014) Selectivity of Big Data. Discussion paper 201411, Statistics Netherlands, The Hague/Heerlen, The Netherlands.

Daas, P.J.H., Puts, M.J., Buelens, B., van den Hurk, P.A.M. (2015) Big Data and Official Statistics. Journal of Official Statistics 31 (2), in press.

R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Zhang, L.-C. (2012), On the accuracy of register-based census employment statistics, Q2012, Athens.