USAGE OF NON-PROBABILITY SAMPLE AND SCRAPED DATA TO ESTIMATE PROPORTIONS

V. Nekrašaitė-Liegė[1,2], A. Čiginas[1,3] and D. Krapavickaitė[2]

[1] Statistics Lithuania, Lithuania
e-mail: Vilma.Nekrasaite-Liege@stat.gov.lt, Andrius.Ciginas@stat.gov.lt

[2] Vilnius Gediminas Technical University, Lithuania
e-mail: Vilma.Nekrasaite-Liege@vilniustech.lt, Danute.Krapavickaite@vilniustech.lt

[3] Vilnius University, Lithuania
e-mail: Andrius.Ciginas@mif.vu.lt

Abstract

An increasing amount of data sources suggests a task to integrate them with the ordinary data sources used in official statistics. One of the problems under the study at Statistics Lithuania is to revise some indicators and to find out if there is room for their accuracy improvement using data from additional sources. The proportion of companies possessing the websites is one such indicator. Traditionally it is estimated using the data of the Information and Communication Technology sample survey.

Information about enterprise website possession is provided also by a private company. However, this data source is updated on a voluntary basis and has some drawbacks: it does not cover all the population, thus the estimator based on this data source should be biased (Tam and Kim, 2018).

Another way to create a list of enterprises owing the websites is to do it by web scrapping (ESSnet Big Data I, ESSnet Big Data II). Following a common methodology, ten potential URLs are found for each enterprise applying a search engine to the population. A logistic regression model is used to estimate the probability, that the selected URL is a website of the particular enterprise. If this probability reaches the fixed threshold, then a conclusion, that the enterprise owns the website, is made. Otherwise, the conclusion is opposite. However, it is known from other research sources, that the accuracy of such an enterprise classification is around 59-89 percent truthful and depends on a search engine, training sample, etc.

Therefore, it may seem that there is no possibility of renouncing the collection of the data on websites through the ICT survey, however, the combination of different sources may lead to more efficient estimators. See Beaumont (2020), Kim and Tam (2021) and Rao (2021) among others.

In this research, the number of methods to integrate auxiliary data obtained from alternative sources with the survey data for bias adjustment is examined. The integration leads to more efficient estimators in comparison with the estimators based only on the survey data. The accuracy measures of the estimators considered are evaluated.

Keywords: Big data, coverage bias, post-stratification, calibration weighting, accuracy estimation.

References

Beaumont, J.-F. (2020). Are probability surveys bound to disappear for the production of official statistics? *Survey Methodology* **46**(1), 1–28.

ESSnet Big Data I. WP2 led by Monica Scannapieco/ISTAT (OBEC) https://ec.europa.eu/eurostat/cros/content/wp2-webscraping-enterprise-characteristics_en

ESSnet Big Data II. WPC led by Galia Stateva/BNSI (OBEC) https://ec.europa.eu/eurostat/cros/content/WPC_Enterprise_characteristics_en

Kim, J.-K. and Tam S.-M. (2021). Data integration by combining big data and survey sample data for finite population inference. *International Statistical Review* **89**(2), 382–401.

Rao, J. N. K. (2021). On making valid inferences by integrating data from surveys and other sources. *Sankhya* **B 83** (1), 242–272.

Tam, S.-M. and Kim J.-K. (2018). Big data, selection bias and ethics – an official statistician's perspective. *Statistical Journal of the IAOS* **34**, 577–588.