Inference for big data assisted by small area methods: an application on SDGs sensitivity of enterprises in Italy

Gaia Bertarelli

Ca' Foscari University of Venice

We propose a new method to estimate the sustainable development goals (SDGs) sensitivity of enterprises in Italy at the provincial level using web-scraping data (a nonprobability sample) because this value is not surveyed by the Italian National Institute of Statistics. The proposed method uses a probability sample to reduce the selection bias of estimates obtained from the nonprobability sample in the context of small area estimation and integrates the nonprobability and probability samples using a double robust estimator that combines (i) propensity weighting to improve the representativeness of the nonprobability sample and (ii) a statistical model to predict the units that are not in the nonprobability sample. A bootstrap procedure for estimating variance is also proposed. To validate the proposed method, Monte Carlo simulation scenarios and an application with real data for e-commerce prevalence were performed. Results show that the proposed method allows the correction of bias from the nonprobability sample while maintaining a good level of estimate reliability.

Co-authors: Francesco Schirripa Spagnolo, Nicola Salvati, Stefano Marchetto, and Monica Pratesi