Small area estimation for big data sources

Francesco Schirripa Spagnolo*[1], Gaia Bertarelli[2]; Stefano Marchetti[1]; Monica Pratesi[1]; Nicola Salvati[2];

[1] University of Pisa
[2] Sant'Anna School of Advanced Studies

Abstract:
Nowadays, the availability of huge amount of data produced by a wide range of new technologies, so called big data, is increasingly. At the same time, their availability to unprecedented spatial detail represents an opportunity in the context of small area estimation to infer some characteristics for very small domains. However, data obtainable from big data sources are often the result of a non-probability sampling process and adjusting for the selection bias is an important practical problem. In this paper, we propose a novel method of reducing the selection bias associated with the big data source in the context of Small Area Estimation (SAE). Our approach is based on data integration and onto the combination of a big data sample and a probability sample. Here we are interested in the estimation of the population mean of a target variable in each small area of interest. We assume the target variable is available from the big data sources, while auxiliary variables present in the big data are also available from survey samples. Because of the selection bias, the sample mean of the target variable calculated using the big data is biased and by incorporating the auxiliary information from an external source, we can reduce the selection bias. In particular, we develop doubly robust estimators by using of small area models with area-specific effects. These models are implemented to obtain the small area estimator from the sample data and the parameters of the propensity score for the big data sample. The proposed estimators have been evaluated in simulation scenarios. Results tend to confirm good performance of our proposal at small area level.

Keywords:
Data integration; Small Area Estimation; Big data; Robust models