Data integration by combining big data and sample survey data for finite population inference

Technological advances have enabled increased access to a growing amount of big data for potential use in producing statistics. Big data can, however, suffer from quality issues, making it challenging to use these data sources to form valid statistical inferences for the population. These issues include statistical bias arising from under-coverage of the population of interest in the dataset, and measurement errors in the variables that are on the data set. We present a regression data integration estimator that can be applied to a probability sample linked to the big data in order to address these issues. This approach can be extended to deal with non-response in the probability sample, and to account for imperfect linkages between records in the probability sample and the big data. The method is applied to a real data example using 2015-2016 Australian Agricultural Census data.

Key words: Calibration weighting; Measurement error; Non-response; Regression estimation; Selection bias.