Model-based procedures for inference from big data sources

Bart Buelens*
Statistics Netherlands, Heerlen, Netherlands – b.buelens@cbs.nl

Joep Burger
Statistics Netherlands, Heerlen, Netherlands – j.burger@cbs.nl

National Statistical Institutes and other organizations get access to a new generation of data sources: big data. Examples include social media messages, mobile phone meta data, traffic flow counts, and odometer readings. As big data sources are large non-probability samples and differ in various aspects from typical administrative registers, methods currently in use at NSIs break down. The question if and how these data can be used to produce official statistics is at the core of the big data research program at Statistics Netherlands. Model-based procedures are considered the most promising family of methods to generalize estimates from big data sources to an intended target population. Predictive data-driven modeling techniques encountered in the fields of data mining and statistical learning provide us with promising alternative tools. A simulation study is conducted using odometer readings of 7.5 million cars. Skewed data sets are generated and used for inference. Pseudo-design-based methods, classical model-based methods and predictive modeling techniques are applied, compared and evaluated. The extent to which modeling approaches are capable of removing sample bias compared to design-based methods is empirically investigated. Model-based methods may allow for unbiased inference in certain scenario's where design-based methods cannot be used.

Keywords: predictive modeling; non-probability sample; pseudo-design-based inference.