Big data, selection bias, and ways to correct for it

Piet Daas*
Statistics Netherlands, Heerlen, The Netherlands – pjh.daas@cbs.nl

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

Abstract

Big data contain traces of human or economic activity that could potentially be used for official statistics. However, big data do not always represent a random sample of the target population, indicating that they are potentially selective. This is a major concern and may result in biased estimates. This paper addresses different uses of big data, discusses the extent to which selectivity is an issue, and proposes methods to handle this.
The three cases in which big data can be used in official statistics production are: i) survey based with big data as an additional source, ii) census based with big data as the single or main source, and iii) big data based with the target population not completely covered.

i) In the first case, the results obtained from survey or admin data essentially lay the foundation. Big data can be used as an additional source of information in two different ways, either as an additional data collection mode or as auxiliary information in a model-based inference approach. An example of the first is the use of scanner data and data collected by web robots as input for the consumer price index. Examples of the second are using big data in the study of the effect of social media sentiment on the consumer confidence survey and the use of mobility data for the study of well-being and other socioeconomic phenomena.
ii) In the second case of using big data, focus is on the coverage of the target population in the source. When the latter is completely included, one can produce a so-called census-based big data statistic. Examples of this are: traffic intensity statistics based on road sensor data, a consumer price index solely based on product prices collected from the web, land-use statistics based on satellite images of a country and waterway transport statistics based on inland Automatic Identification System transponder data from ships. The first and the second example have actually been produced.
iii) In the third case, when the target population is not completely included in big data, selectivity is a serious issue. Finding ways to deal with it becomes the most important question. Examples of studies in which this is obviously the case are: a social media based unsafety monitor, a social media based sentiment index, day time population statistics based on mobile phone data and tourism statistics based on mobile phone data. In the second and third case attempts were made to correct for the missing part of the population. However, even though the findings are or may still be biased, they also reveal the potential of using big data in these specific areas. However, before correcting for selectivity, there are challenges that must be dealt with first. These are: the fact that most big data sources are event (and not unit) oriented, and that many of these sources do not contain directly available (background) characteristics of the units included. The latter has stimulated the need to find ways to derive them from the data, which is demonstrated with examples.

This paper elaborates on the current view on these matters at Statistics Netherlands, examples are provided for each of the above mentioned cases, and recommendations are made including some discussion points for future research.

Keywords: representativeness, selection bias, feature extraction, data science.