Title: On big data based statistical inference
Authors: Piet Daas, Marco Puts and Robert Renssen (Statistics Netherlands)

Big data is a hot topic and many scholars and data scientists mention they are studying it. Usually this is done with a particular application in mind and not from a generic (big data) viewpoint. For big data, the latter is essential as it is becoming more and more apparent that -to produce big data based statistics- a new way of working is required [1]. Henceforth, there is an urge to have a general, fundamental statistical framework to describe phenomena hidden in big data. If such a fundament would be available, it will guide researchers on how they can make optimal use of the data in such sources. Since the authors of this abstract are statisticians, we note an analogy with sampling theory here. The big question for us is: “What would the big data analogue of sampling theory look like?” Since big data is usually not created with a particular statistical purpose in mind, it is obvious that the use of big data is of a secondary nature. This makes its application for statistics a derived one; similar to the use of administrative data. This use may affect the units generating the data in the source and the (value of the) variables measured. Both are prone to error, since i) the units in the source may not be identical to those of the target population and ii) the variable included in the source may not be the variable the researcher had in mind. So suppose a researcher wants to determine the mean of the unknown variable μ of a target population. In the case of an administrative source, covering the whole population, he can operationalize this by determining y for $i = 1, \dots, N_{pop}$; with y_i the values for each of the N units in the population. When a big data source is available with a variable z, he can operationalize this by determining z for the big data population for $j = 1, \dots, M_{BD}$. In both cases it is assumed that y and z are approximations of μ. Hence the mean of μ is approximated by:

$$\bar{\mu} \cong \hat{\bar{y}} = \frac{1}{N} \sum_{i=1}^{N_{pop}} y_i \quad (1), \qquad\qquad \bar{\mu} \cong \hat{\bar{z}} = \frac{1}{M_{BD}} \sum_{j=1}^{M_{BD}} z_j \quad (2)$$

In formula (2) it is assumed that the big data source includes a huge part of the units of the target population, i.e. the units missing can be ignored. The more z is related to y and the more the big data population M is equivalent to the target population N, the more likely the series of z and y correlate and cointegrate. The latter refers to the co-movement of both series over time. The formulas indicate two important issues. The first is related to the populations compared. A difference in the composition of the target population and the population generating the data in the big data source may result in a biased estimate. The second is related to the variables compared. Both variables are approximations of the original concept μ. Because of this also a bias may be introduced. However, if the bias of both variables remains constant over time, this does not have to be a huge problem. In the latter case, both series would nicely correlate and probably also cointegrate. An interesting example of this are the time series of consumer confidence (survey based) and social media sentiment (big data based) in the Netherlands. Both series show a remarkable high correlation; r = 0.9 [2, 3]. This unexpected finding seems to suggest that large scale data sets may suffer less from the issues identified above. They are still important however, as studying a smaller subset of big data, for instance a local area, seriously reduces its size and may cause to re-surface these effects. It is important to notice that each data source, whether it is a survey sample, admin or big data source, provides its own way to measure a particular concept. Each is an approximation of the target variable μ. The presentation or poster will expand on this work and include examples.

[1] WP8 (2017) Results of Work Package 8, on Methodology, Quality and IT, of the ESSnet Big Data project.

[2] Daas, P.J.H., Puts, M.J.H. (2014) Social Media Sentiment and Consumer Confidence. European Central Bank Statistics Paper Series No. 5, Frankfurt, Germany.

[3] Van den Brakel, J., Söhler, E., Daas, P., Buelens, B. (2017) Social media as a data source for official statistics; the Dutch Consumer Confidence Index. *Survey Methodology,* accepted for publication.