Review of: "Big Data, Granger Causality Analysis, and the Undecidability Property of Neuroimaging"

Aleksandra Matuszewska-Janica[1]

1 Warsaw University of Life Sciences

Potential competing interests: The author(s) declared that no potential competing interests exist.

The article concerns approaches to the analysis of large datasets from neuroimaging area. The Author clearly points out that an atheoretical approach to data analysis in this area is not defensible. Given the expanding datasets available and the increasing opportunities for processing these data, discussions in this area are important and valuable. It is worth noting that the EOT problem in many fields (e.g. economics) has long since been resolved in favour of reconciling atheoretical and theoretical approaches. I think that much in this regard is also explained by Breinman (2001).

Several elements in the article are open to discussion and, in my view, require some clarification. In general, the concept of big data is not clearly defined in the literature, as Ward and Backer (2013) highlight. Therefore, it would be helpful if the Author clearly defines the conceptual area in which he is operating (with the reference to the source of these definitions). Most researchers tend to adopt an intuitive definition of big data as a large or complex dataset that is too problematic to analyse using 'traditional' software. In contrast, what the author calls the big data approach is more closely associated with data exploring or data mining.

There are several inaccuracies when describing Granger causality (Section 1). The Author writes: “A ‘variable’ is a stochastic time series that is related to itself, typically, using time lag operators”. Time series is usually defined as “a sequence of data points that occur in successive order over some period of time”. What the author refers to as 'time series that is related to itself' refers to autocorrelation in a time series. Such a process can be described by a family of autoregressive models. Generalizing this to a set of multiple variables leads to a representation in the form of vector autoregression models (VAR or MVAR as Kaminski et al. 2001 specify).

The Author defines Granger causality as follows “Given the parameters of the Granger causality hypothesis test,$X_1(t)$ “G-causes” $X_2(t)$ if E_1 is minimized by including the X_2 term in the top equation more than compared to when the term is absent”. In fact, what is at issue here is either a significant reduction in prediction error (see e.g. Charemza and Deadman 1992) or a significant reduction of the variance of prediction error as report Kaminski et al. (2001).

I do not understand the reasoning behind introducing the lag operator (L) if it is not used in the rest of the work. I think it would also be useful to add an explanation of why, in the case cited, the Geweke approach is preferable to the one originally proposed by Granger (1969). This is obvious to those familiar with the method; however, it would be a nod to those new to these methods.

Breiman L., (2001), Statistical Modeling: The Two Cultures, Statistical Science, Vol. 16, No. 3 (Aug., 2001), pp. 199-215,

http://www.jstor.org/stable/2676681

Charemza, W. W., Deadman, D.F., (1997), New Directions In Econometric Practice, Second Edition, Edward Elgar Publishing.

Ward, J. S., & Barker, A. (2013). Undefined by data: a survey of big data definitions.*arXiv preprint arXiv:1309.5821*.