Digital proxies

written by Jeremy Huggett | 28/11/2017

The US Department of Immigration and Customs Enforcement (ICE) is apparently seeking to employ 'big data' methods for automating their assessment of visa applications in pursuit of meeting Trump's calls for 'extreme vetting' (e.g. Joseph 2017, Joseph and Lipp 2017, and see also). A crucial problem with the proposals has been flagged in a letter to the Acting Secretary of Homeland Security by a group of scientists, engineers and others with experience in machine learning, data mining etc.. Specifically, they point to the problem that algorithms developed to detect 'persons of interest' could arbitrarily select groups while at the same time appearing to be objective. We've already seen this stereotyping and discrimination being embedded in other applications, inadvertently for the most part, and the risk is the same in this case. The reason provided in the letter is simple:

> "Inevitably, because these characteristics are difficult (if not impossible) to define and measure, any algorithm will depend on 'proxies' that are more easily observed and may bear little or no relationship to the characteristics of interest" (Abelson et al 2017)

Leaving the important political questions raised by the ICE proposals and the chilling effect on potential visitors to the US aside, the challenges and risks of using proxies highlighted by this debate are also relevant to archaeology. After all, if we are interested in understanding the meanings behind the physical material culture that archaeologists habitually deal with, we are forced to use proxies (though we don't often call them that) as a means of interpreting the physical evidence in more than simple descriptive terms – for instance, proxies are the means by which we move from the shape, form, and distribution of pottery to modes of exchange. For example, Bjørnar Olsen has written about how we primarily seem concerned with material culture as a 'stand-in' or proxy for something else (social, political, cultural, ideological etc.) (2010, 3).

Like the big data techniques at issue in the ICE case, we use proxies in archaeology when what we are interested in cannot be seen or measured directly, in the belief that such proxies allow us to access the immaterial processes behind the tangible evidence we have to hand. So, for example, we have become increasingly accustomed to using the idea of visibility as a proxy for knowledge in GIS, or friction as a proxy for accessibility, artefact density for levels of human activity, radiocarbon plots for prehistoric occupation, tombs as indicators of settlement, the substitution of one form of material culture for another as a measure of population replacement, material culture traits as proxies for social identity and/or group membership, or trade and exchange, and so on. For instance, Thomas Whitley employed a set of spatial proxies in his GIS analyses of cognition (2004). More recently, Scott Gallimore (2017) has written about the use of proxies in studying food surpluses in Roman Crete, using amphorae as proxies for their contents (wine rather than olive oil in

this case).

Nor is this simply a feature of the more ‘humanistic’ or ‘social’ approaches within archaeology. For instance, proxies are used extensively in environmental archaeology and climate reconstruction. Since detailed measurements have only been captured over the last century or so, variability and change through time is approached through proxies such as pollen, cores, borehole data, tree rings, corals, sediments, etc. (e.g. Mann 2002). Multi-proxy analysis is frequently used, whereby several proxies are employed in combination or as checks and balances against each other. Much the same is seen in GIS analyses – for instance, locational models are frequently the result of multi-proxy analysis, where elevation, slope, aspect, proximity etc. are used as measurable variables which can give rise to a model of preferential location, for example. Indeed, Tallavaara *et al* (2014, 137) usefully identify three categories of proxy:

- proxies tracking temporal changes in the amount of archaeological material (such as frequency distributions);
- proxies that are not dependent on the amount of archaeological material (such as proportional measures);
- proxies that are independent of the archaeological record (e.g. genetic or simulation studies).

One might assume that a multi-proxy analysis employing proxies derived from each of these categories should provide powerful complementary evidence confirming or negating the conclusions being drawn.

None of this is to suggest that because they are commonplace, proxies are simple and straightforward in archaeology. Debates concerning their use would indicate otherwise – for instance, Tallavaara *et al* 2014, Timpson *et al* 2015, and Smith 2016 are examples of an extensive debate over a number of years concerning radiocarbon dates as proxies for population).

What is interesting, though, is to consider the use of proxies in the light of the growing interest in ‘big data’ approaches in archaeology, and Anderson’s famous declaration of the “end of theory” in which “correlation is enough” (Anderson, 2008). And course, our archaeological interpretations are based upon correlation (or otherwise) between our selected proxies and the features of interest. But what big data methods offer is the ability to identify proxies without the need for hypothesising a relationship between these features and our proxy(ies) – we simply apply our computational tools to all of our data and let them identify the most appropriate proxy for us. Hence “No longer do we necessarily require a valid substantive hypothesis about a phenomenon to begin to understand our world” (Mayer-Schönberger and Cukier 2013, 55) – our analysis becomes data-driven instead of hypothesis-driven, although unfortunately for Mayer-Schönberger and Cukier, their primary example of Google Flu Trends has since been discredited (e.g. Fung 2014).

However, it is also interesting to see how some of the arguments supporting the use of proxies within archaeology can also be used to support the use of big data techniques. For instance, Timpson *et al* (2015) suggest that we can perhaps set aside concerns over biases in the data, that it is not the case that attempting to remove the biases necessarily improves the quality and hence reliability of the resulting inferences (2015, 200-201). They offer three reasons why this might be the case:

> "Firstly, archaeological data are often frustratingly sparse, and this causes a large sampling error that can easily dwarf the effects of particular biases. Secondly, all data are subject to many different biases. By using the broadest possible inclusion criteria from multiple sources, the Law of Large Numbers predicts that the combination of many different biases will approach a random error. Thirdly, dirty data will have the effect of hiding (adding noise to) any true underlying pattern. This will certainly make it harder to detect what is really going on, but this has the desirable effect of making the null hypothesis harder to reject, thus making the statistical test conservative." (2015, 201).

The problem remains, though, that whether digital or not, our proxies must bear some relation to our items of interest and we cannot assume that, just because there is a correlation, it is necessarily a meaningful one. For example, Timpson *et al* argue that all proxies contain some information about the quantity of interest and illustrate their claim with an example of a correlation between ice cream sales and the murder rate (since both turn out to be affected by the independent proxy of temperature) and suggest that if the correlation is strong it may therefore be an excellent proxy (2015, 200). That's as may be, but does it make any sense? Would not the more logical approach be to connect temperature with the murder rate or ice cream sales instead? And isn't the melting pot of big data even more likely to connect similarly illogical proxies and compound the problem by doing so invisibly?

References

Anderson, C. (2008), 'The end of theory', *Wired magazine* 16(7), 16–07.
https://www.wired.com/2008/06/pb-theory/

Fung, K. 2014 'Google Flu Trends' Failure Shows Good Data > Big Data', *Harvard Business Review Digital Articles* March 25, 2014, 2-4.
https://hbr.org/2014/03/google-flu-trends-failure-shows-good-data-big-data

Gallimore, S. 2017 'Food surplus and archaeological proxies: a case study', *World Archaeology* 49 (1), 138-50. https://doi.org/10.1080/00438243.2016.1259582

Joseph, G. 2017 'Extreme Digital Vetting of Visitors to the U.S. Moves Forward Under a New Name', *ProPublica* (22/11/17)
https://www.propublica.org/article/extreme-digital-vetting-of-visitors-to-the-u-s-moves-forward-under-a-new-name

Joseph, G. and Lipp, K. 2017 'How ICE Is Using Big Data to Carry Out Trump's Anti-Immigrant Crusade', *Splinter* (8/11/17)
https://splinternews.com/how-ice-is-using-big-data-to-carry-out-trumps-anti-immi-1797745578

Mann, M. 2002 'The Value of Multiple Proxies', *Science* 297, 1481-1482.
http://www.meteo.psu.edu/holocene/public_html/shared/articles/MannPersp2002.pdf

Mayer-Schönberger V. and Cukier K. 2013 *Big Data: A Revolution that Will Change How We Live, Work and Think.* London, John Murray.

Olsen, B. 2010 *In Defense of Things: Archaeology and the Ontology of Objects*, Altamira Press.

Smith, M. 2016 'The use of summed-probability plots of radiocarbon data in archaeology', *Archaeology in Oceania* 51, 214-219. https://doi.org/10.1002/arco.5094

Tallavaara, M., Pesonen, P., Oinonen, M. and Seppä, H. 2014 'The Mere Possibility of Biases Does Not Invalidate Archaeological Population Proxies – Response to Teemu Mökkönen', *Fennoscandia Archaeologica* 31, 135-140. http://www.sarks.fi/fa/PDF/FA31_135.pdf

Timpson, A., Manning, K., and Shennan, S. 2015 'Inferential mistakes in population proxies: A response to Torfing's "Neolithic population and summed probability distribution of ^{14}C-dates"', *Journal of Archaeological Science* 63, 199-202. https://doi.org/10.1016/j.jas.2015.08.018 (authors' accepted version available via http://discovery.ucl.ac.uk/1470834/3/manning_discovery_template.pdf)

Whitley, T. 2004 'Spatial Variables as Proxies for Modelling Cognition and Decision-Making in Archaeological Settings: A Theoretical Perspective', *Internet Archaeology* 16. https://doi.org/10.11141/ia.16.3