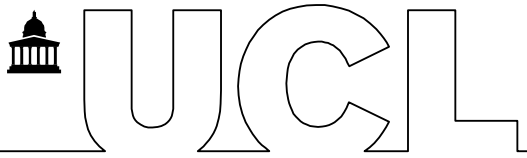

Timpson, A; Manning, K; Shennan, S; (2015) Inferential mistakes in population proxies: A response to Torfing's "Neolithic population and summed probability distribution of ¹⁴C-dates". **Journal of Archaeological Science** 10.1016/j.jas.2015.08.018. (In press). Downloaded from UCL Discovery: http://discovery.ucl.ac.uk/ 1470834

ARTICLE

Inferential mistakes in population proxies: A response to Torfing's "Neolithic population and summed probability distribution of ^{14}C-dates"

Adrian Timpson[1], Katie Manning[2], Stephen Shennan[2]
[1] Department of Genetics, Evolution and Environment, University College London, Darwin Building, Gower Street, London WC1H 0PY
[2] Institute of Archaeology, University College London, 31-34 Gordon Square, London WC1H 0PY
Corresponding author: Adrian Timpson Email: a.timpson@ucl.ac.uk

Abstract
In his paper *"Neolithic population and summed probability distribution of ^{14}C-dates"* Torfing opposes the widely held principle originally proposed by Rick (1987) that variation through time in the amount of archaeological material discovered in a region will reflect variation in the size of that local human population. His argument illustrates a persistent divide in archaeology between analytical and descriptive approaches when using proxies for past population size. We critically evaluate the numerous inferential mistakes he makes, showing that his conclusion is unjustified.

Introduction

In his paper *"Neolithic population and summed probability distribution of ^{14}C-dates"* Torfing (in press) opposes the widely held principle originally proposed by Rick (1987) that variation through time in the amount of archaeological material discovered in a region will reflect variation in the size of that local human population. Torfing uses a radiocarbon dataset covering the Jutland peninsula to produce a Summed Probability Distribution (SPD) to make comparisons with studies by Hinz et al 2012 and Shennan et al 2013, in order to explore the effects of three particular human related biases on the SPD. Specifically, he shows that the shape of the SPD changes when the following subsets of samples are excluded: (1) samples from shell middens, (2) samples associated with visible monuments, and (3) samples from research conducted by one particular archaeologist, S. H. Andersen. Torfing then argues that these differences in shape demonstrate that the SPD is detrimentally subject to three biases: changes in subsistence strategy, changes in ritual actions, and modern research strategies. This, he concludes "*…induces a general doubt about the validity of sum probability distributions as a population proxy*".

Torfing's opposition illustrates a persistent divide in archaeology between analytical and descriptive approaches. His arguments are representative of a long-standing attitude among

many archaeologists that, unless we have complete knowledge of all the factors that might possibly affect the record available to us, which of course we never will, then we cannot say anything at all.

Descriptive approaches have their place, but if archaeology is to be more than mere catalogues of material then we need to go beyond them and address the factors accounting for variation in that material, as Binford (1962) long ago pointed out. There are good reasons to believe that changes in population size and density may be among those factors, and indeed they have often been invoked, but the study of past population size is unquestionably an objective and quantitative issue requiring an analytical approach. SPDs have evolved out of the rudimentary approach of counting the number of sites per archaeologically-defined period and have, in recent years, become an increasingly widespread method in archaeology and more broadly in Quaternary science (Gamble et al. 2005; Johnstone et al. 2006; Buchanan et al. 2008; Collard et al. 2010; Johnson and Brook 2011; Manning & Timpson 2014). Many early studies presented the SPD as the final product, as does Torfing, but in doing so he ignores the key contribution made in Shennan *et al*. (2013), which was the application of the scientific method of *explicitly testing a null hypothesis* – in our case that European populations exhibited a steady (exponential) growth from the Mesolithic onwards – rejecting the null only if the SPD curve significantly differs from expectation. This study showed that the shape of an SPD is merely descriptive unless it is rigorously tested in light of potential biases and uncertainties, such as the calibration process, ascertainment bias and sample size.

Ultimately, Torfing's failure to apply this scientific approach is the cause of his erroneous conclusion. His subjective criticisms, lacking in any formal hypothesis testing, only serve to stagnate the discipline. It is essential to critically evaluate and address the conservative limits of Rick's basic assumptions when using SPDs, but this needs to be done within a scientific framework if we hope to move current methods forward. Building on other constructive contributions to the method (e.g. Peros et al. 2010; Johnson and Brook 2011; Williams 2012), the objective of Shennan *et al*. (2013) and Timpson et *al*. (2014) was to do precisely that by specifically addressing recent criticisms such as taphonomic bias (Surovell *et al*. 2009); spurious calibration effects (Chiverrell *et al*. 2011; Bleicher 2013) and temporal resolution (Contreras and Meadows 2014).

Rather than construct a lengthy comprehensive rebuttal of every issue, the following deals with a few of the more general inferential mistakes that Torfing has made that we consider relevant to the broader field.

Inferential mistakes

Missing the whole point of a proxy

The presence of biases in radiocarbon sampling is undeniable. Rick himself identified many sources of error that contribute both noise and other undesirable non-random signals that prevent a perfect correlation with the human population level. This issue is unique to neither radiocarbon samples, nor SPDs, but is fundamental to the nature of any proxy. A proxy is used when the quantity we are interested in cannot be measured directly. As such, all proxies will contain information from other processes, resulting in a less than perfect correlation with the quantity of interest. Therefore it is unhelpful for Torfing to categorise a proxy as either valid or invalid, since all proxies contain *some* information about the quantity of interest. Consider for example, that ice cream sales correlate rather nicely with the

murder rate. It turns out this is because both are affected by temperature, along with the shark-attack rate and sandal-wearing. So are ice cream sales a proxy for the murder rate? If the correlation is strong then this may indeed provide an excellent proxy, even in this peculiar example where neither is the direct cause of the other.
A careful evaluation of a particular bias can enable us to incorporate other relevant data into a statistical model to improve how well it correlates with the quantity of interest. Even if a model cannot be improved, the error can be estimated, to provide a confidence range. Further work in this direction would make a valuable contribution to the archaeological toolkit when evaluating past population proxies. Unfortunately, Torfing does not take this progressive path. On the contrary he repeatedly dismisses this as *"…a hopeless endeavour*", since *"…no dataset will be large enough to overcome this, as the bias is naturally built into the data."* and *"…we cannot with any accuracy determine how to compensate for this systematic bias"*.

Of course the subsets look different, but this doesn't demonstrate a bias

By dividing the dataset into smaller subsets, Torfing substantially reduces the sample size (from 463 dates to 162 dates in dataset 1, and 127 to 56 dates in dataset 2). The effect is a massive increase in sampling error, which will change the shape of the SPD. This is important since we should expect similar changes even if no biases exist at all, and even if the subsets are randomly sampled from the true underlying distribution. Therefore Torfing's inference that the change in shape is a demonstration of a bias is completely unjustified. Of course we are not arguing that biases do not exist – on the contrary, we are advocating methodological improvements to deal with them. But clearly Torfing has failed to untangle the effects of biases from the inevitable and substantial differences expected merely from sampling. He has made no attempt to quantify the differences in shapes, let alone test if these differences are statistically significant. One reader might subjectively interpret these differences as substantial, whilst another might consider them small enough (given the decrease in sample size) to be meaningless. It is unacceptable that Torfing fails to test for sampling effects given that statistically rigorous methods which account for precisely this problem are used in the very papers that he is criticising.(Shennan *et al.* 2013, Timpson *et al*. 2014).

No sample is a perfect representation of the population, but larger samples from a broader inclusion strategy are fairer.

In section 3.3 Torfing argues that one particular researcher S.H Andersen had certain research biases, and therefore removes radiocarbon dates obtained by this researcher. This is a fundamentally flawed approach for several reasons.

Biases are ubiquitous at the scale of individual researchers. Torfing's approach of excluding data from a researcher with a specific research interest has the hugely detrimental effect of reducing the sample size. Indeed, following this approach would logically result in the exclusion of data from every researcher until nothing remains. When we suspect a bias in some data (or that they are otherwise unreliable/erroneous) it is tempting to assume that their exclusion must improve the overall quality, and therefore the reliability of the inferences drawn. Surprisingly this is rarely the case for archaeological data. There are three reasons for this. 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.

Moreover, the effects of individual researcher bias have obvious limits. Even with the most constrained and specific research interests, no researcher can know the date of a sample before he/she sends it to the laboratory. After the sample has been processed, however, it will contribute to the radiocarbon record even if it dates from a period of no interest to the researcher. The radiocarbon record therefore has a natural tendency to mitigate the effect of individual research interests.

Obtaining a different SPD by removing a subset does not undermine the SPD as a population proxy.

Torfing suggests that the monumentality of some Neolithic structures makes them an invalid proxy for human activity, and that instead only settlement data should be used. But, are we therefore supposed to assume no relationship between the number of tombs and the number of communities that built them and buried their dead there? If the record only consisted of burials would Torfing conclude that it was devoid of settlement? One of Rick's basic assumptions was that the amount of human activity in a given region acts as a useful proxy for the number of people undertaking those activities. Therefore, by excluding the remains of certain activities from the landscape (especially those that represent an investment of labour and human resources) only serves to bias the representation of human activity. Thus, to take a relevant example, the pattern produced by looking only at the settlement data, as opposed to all the evidence for human activity, does not correspond to the pollen evidence for 4th millennium BCE human impact from the adjacent area of Schleswig-Holstein, which has a similar sequence to Jutland (Hinz et al. 2012, Feeser et al. 2012), nor more generally to the pollen evidence for northern Europe as a whole (Nielsen et al 2012). In fact, in Schleswig-Holstein Feeser and Furholt (2014) show that for the TRB period the relationship between economic activity evidenced in pollen diagrams and ritual activity indicated by megalithic tomb construction has an R^2 value of 0.8, demonstrating that they are extremely highly correlated and that tombs can be taken as an excellent indicator of settlement activity.

Even if Torfing had gone to the trouble of statistically testing his SPDs and found significant differences, would this support his conclusion that the overall SPD does not correlate with the population? We argue that this inference is completely unjustified. A simple thought experiment demonstrates the problem. Let us assume a hypothetical SPD for a particular region that shows a constant level through time, with the exception of one large peak across a 300 year period which is caused entirely by a subset of samples that are associated with visible monuments (i.e. repeating Torfing's approach of removing the monument samples produces a flat SPD). Certainly we can infer that the monuments were likely only built during this 300 year period, but can we also infer that the human population was constant through time and the radiocarbon record was over-represented by this subset? Perhaps the population did increase during the 300 year period, precisely in proportion to the increased evidence from the monuments. Alternatively the monument samples might under-represent the human population and in fact there was a much greater increase in population than the

SPD suggests. Or perhaps it over represents the population. This problem applies to the exclusion of any individual category of samples or material type that has not been constant through time. For example, future archaeologists may notice the sequential rise and fall of record players, audio cassette, CDs and iPods through time, and SPDs constructed for each separately would significantly differ from each other. However the individual SPDs would only describe the popularity of those objects, and tell us little about the actual population size. In contrast we can expect a combined SPD of all objects to act as a better proxy for population size through time as it accumulates more information about human activity. This illustrates why the broadest inclusion criteria of all samples with an anthropogenic association is the most informative strategy.

Summary

The use of SPD analysis has become increasingly widespread and sophisticated, as more researchers contribute intelligent methodological advancements. In contrast, Torfing's claims entirely lack statistical support and are riddled with inferential mistakes. He provides descriptive arguments for the presence of three specific biases in order to raise doubts that SPDs are a valid population proxy, but makes no efforts to either quantify these biases, or hypothesis test his claim. As such Torfing misses the whole point of the scientific method. Science does not progress through the accumulation of doubts until we grind to a halt in a bog of ignorance. Instead it stacks the cards in favour of the best existing hypothesis, then conservatively tests new ideas using the most critical tools available. Occasionally the *status quo* can be rejected, and this represents a tentative new step towards a better place. We don't share Torfing's nihilistic view that all progress in this area is doomed, and find no justification for his conclusion that SPDs are an invalid population proxy. Nevertheless we recognise that (as is the case with any proxy for a quantity that cannot be measured directly) it is important to continue in our critical evaluation of *how well* SPDs correlate with true past population, and how detrimental signals from biases or other spurious sources can be better managed.

Acknowledgements

We are grateful to Mike O'Brien for his helpful comments and suggestions, and to the European Research Council for Advanced Grant #249390, 'Cultural Evolution of Neolithic Europe' that has supported this work.

References

Binford, L.R. 1962. Archaeology as anthropology. *American Antiquity* 28, 217-225.

Bleicher, N. 2013. Summed radiocarbon probability density functions cannot prove solar forcing of Central European lake-level changes. Holocene 23(5): 755-765.

Buchanan, B., Collard, M. & Edinborough, K. 2008. Palaeo-Indian Demography and the Extraterrestrial Impact Hypothesis. *Proceedings of the National Academy of Sciences* 105: 11651–11654.

Chiverrell, R.C., Thorndycraft, V.R. & Hoffmann, T.O. 2011. Journal of Quaternary Science 26(1): 76-85.

Collard, M. Edinborough, K., Shennan, S. & Thomas, M.G. 2010. Radiocarbon evidence indicates that migrants introduced farming to Britain *Journal of Archaeological Science* 37 (4): 866–870.

Contreras, D.A & Meadows, J. 2014. Summed radiocarbon calibrations as a population proxy: a critical evaluation using a realistic simulation approach. *Journal of Archaeological Science* 52: 591-608. doi:10.1016/j.jas.2014.05.030

Feeser, I., Dörfler, W., Averdieck, F-R. & Wiethold, J. 2012. New insight into regional and local land-use and vegetation patterns in eastern Schleswig-Holstein during the Neolithic. In: J. Müller (ed.): *Frühe Monumentalität und soziale Differenzierung*, band 2: 159-190. Bonn: Habelt.

Feeser, I., & M. Furholt 2014. Ritual and economic activity during the Neolithic in Schleswig-Holstein, northern Germany: an approach to combine archaeological and palynological evidence. Journal of Archaeological Science 51: 126–134. doi:10.1016/j.jas.2013.01.021

Gamble, C., Davies, W., Pettitt, P., Hazelwood, M. & Richards, M. 2005. The archaeological and genetic foundations of the European population during the late glacial implications for 'agricultural thinking'. *Cambridge Archaeological Journal* 15: 193-223.

Hinz, M., Feeser, I., Sjögren, K. G., & Müller, J. 2012. Demography and the intensity of cultural activities: an evaluation of Funnel Beaker Societies (4200–2800 cal BC). *Journal of Archaeological Science* 39, 3331-3340.

Johnson, C.N & Brook, B.W. 2011. Reconstructing the dynamics of ancient human populations from radiocarbon dates: 10 000 years of population growth in Australia. *Proceedings of the National Academy of Sciences* 278: 3748-3754; DOI: 10.1098/rspb.2011.0343

Johnstone, E, Macklin, M.G. & Lewin, J. 2006. The development and application of a database of radiocarbon-dated Holocene fluvial deposits in Great Britain. *Catena* 66: 14–23.

Manning, K. & Timpson, A. 2014. The demographic response to Holocene climate change in the Sahara. Quaternary Science reviews 101: 28-35. doi:10.1016/j.quascirev.2014.07.003

Nielsen, A.B., T. Giesecke, M. Theuerkauf, I. Feeser, K.-E. Behre, H.-J. Beug, S.-H. Chen, J. Christiansen, W. Dörfler, E. Endtmann, S. Jahns, P. de Klerk, N. Kühl, M. Latalowa, B.V. Odgaard, P. Rasmussenm, J.R. Stockholm, R. Voigt, J. Wiethold, S. Wolters 2012. Quantitative reconstructions of changes in regional openness in north-central Europe reveal new insights into old questions. *Quaternary Science Reviews* 47, 131-149.

Peros, M. C., Munoz, S. E., Gajewski, K., & Viau, A. E. 2010. Prehistoric demography of North America inferred from radiocarbon data. Journal of Archaeological Science, 37, 656–664.

Rick, J. W. 1987. Dates as data: an examination of the Peruvian pre-ceramic radiocarbon record. *American Antiquity*, 52, 55–73.

Shennan, S., S.S. Downey, A. Timpson, K. Edinborough, S. Colledge, T. Kerig, K. Manning & M.G. Thomas 2013. Regional population collapse followed initial agriculture booms in mid-Holocene Europe. *Nature Communications* 4:2486. DOI: 10.1038/ncomms3486.

Surovell, T. A., Byrd Finley, J., Smith, G. M., Brantingham, P. J., & Kelly, R. 2009. Correcting temporal frequency distributions for taphonomic bias. Journal of Archaeological Science, 36, 1715–1724.

Timpson, A., S. Colledge, E. Crema, K. Edinborough, T. Kerig, K. Manning, M.G. Thomas and S. Shennan 2014. Reconstructing regional demographies of the European Neolithic using radiocarbon dates: a new case-study using an improved method. *Journal of Archaeological Science* 52, 549-557.

Torfing, T. In press. Neolithic population and summed probability distribution of 14C-dates. *Journal of Archaeological Science*, Available online 16 June 2015. Doi:10.1016/j.jas.2015.06.004.

Williams, A.N. 2012. The use of summed radiocarbon probability distributions in archaeology: a review of methods. Journal of Archaeological Science 39, 3:578-589. doi:10.1016/j.jas.2011.07.014