
Article:
Buck, C.E. and Meson, B. (2015) On being a good Bayesian. World Archaeology, 47 (4). 567 - 584. ISSN 0043-8243

https://doi.org/10.1080/00438243.2015.1053977

On Being a *Good* Bayesian

Caitlin E. Buck[1] and Bo Meson

[1]School of Mathematics and Statistics, University of Sheffield, UK
c.e.buck@sheffield.ac.uk

Accepted for publication in World Archaeology doi:10.1080/00438243.2015.1053977

Abstract

Bayesianism is fast becoming the dominant paradigm in archaeological chronology construction. This paradigm shift has been brought about in large part by widespread access to tailored computer software which provides users with powerful tools for complex statistical inference with little need to learn about statistical modelling or computer programming. As a result, we run the risk that such software will be reduced to the status of black-boxes. This would be a dangerous position for our community since *good*, principled, use of Bayesian methods requires mindfulness when selecting the initial model, defining prior information, checking the reliability and sensitivity of the software runs and interpreting the results obtained. In this paper, we provide users with a brief review of the nature of the care required and offer some comments and suggestions to help ensure that our community continues to be respected for its philosophically rigorous scientific approach.

1 Introduction and Background

There is a plethora of books on the market that propound the Bayesian approach to reasoning and inference. The best are very general and take rigorous philosophical (e.g. Howson and Urbach, 1993) and/or mathematical (e.g. Bernardo and Smith, 1994) standpoints, making them hard for those without appropriate training to fully comprehend. The more subject-specific volumes (e.g. Buck et al., 1996; Koop et al., 2007; Kéry and Schaub, 2012; Kaplan, 2014) are more approachable but tend to provide the philosophical and theoretical ideas in separate sections or chapters from those on the applications and, hence, the former are often overlooked. This observation is the starting point for an excellent book by Mayo and Spanos (2010), from the introduction to which the following quote is taken:

> "Methodological discussions in science have become increasingly common since the 1990s... in areas most faced with limited data, error, and noise... To varying degrees, such work may allude to philosophies of theory testing and theory change and philosophies of confirmation and testing (e.g., Popper, Carnap, Kuhn, Lakatos, Mill, Peirce, Fisher, Neyman-Pearson, and Bayesian statistics). However, the different philosophical "schools" tend to be regarded as static systems whose connections to the day-to-day questions about how

> to obtain reliable knowledge are largely metaphorical... The unintended consequence is that the influence of philosophy of science on methodological practice has been largely negative."

Given the above, and the fact that both generic (Lunn et al., 2000, 2009) and problem-specific (e.g. Buck et al., 1999; Haslett and Parnell, 2008; Ramsey, 2009; Blaauw and Christen, 2011) user-friendly software for Bayesian inference is now available, it is very easy for archaeologists and other researchers to use Bayesian methods without appreciating the important underlying philosophical and mathematical assumptions on which they are based. So much so that several so-called "Bayesian revolutions" have taken place in recent years (Brooks, 2003; Beaumont and Rannala, 2004; Shultz, 2007; Wilkinson, 2007; Kruschke et al., 2012; Drummond et al., 2012; Bouckaert et al., 2014); including the one known as the "third radiocarbon revolution" (Bayliss, 2009), with little associated discussion amongst the protagonists of their underlying presumptions.

This risks rendering such endeavours, at best, technically ill-founded – but nonetheless publishable and useful for the user-community – and, at worst, fundamentally scientifically flawed and misleading. Since Bayesian methods currently feature so prominently in the archaeological literature, particularly that relating to the chronology and prehistory of Europe (e.g. Buck et al., 1992; Bayliss et al., 2007a; Manning, 2007; Whittle and Bayliss, 2007; Finkelstein and Piasetzky, 2010; Whittle et al., 2011; Smyth, 2013; Wicks et al., 2014), and since so few of these authors discuss why they have taken a Bayesian approach or, indeed, why they themselves take a Bayesian standpoint on the philosophy of science, now seems a good time to pause and reflect.

To describe oneself or the methods one uses as Bayesian is to make some very specific claims about the way one approaches problems, the tools one uses for tackling them and the way one interprets the results one obtains. One can use Bayesian software simply because one is satisfied by the results one obtains, but this is not sufficient to describe oneself as Bayesian. Nor is it sufficient to describe the results obtained as Bayesian, unless one both uses the software in a rigorous manner and does so in keeping with the underlying tenets of Bayesian theory. It should be noted here that pragmatic use of theory-dependent, degrees of belief in scientific reasoning is not sufficient to justify the label Bayesian. In a similar vein, the idealistic use of theory-independent, 'absolute facts' (often coupled with the casual use of the term 'proof') does not justify the term objective — even in the sense of recognising one's biases, let alone in the more problematic sense of studying 'absolute reality'. The notion of objectivity is further tarnished by the objection (e.g. Haraway, 1988) that it has traditionally allowed the user of the term to distance themselves from any ethical dimension touched on by their research.

In this paper we explore what it means to apply the term Bayesian to a person or a methodology and ask: what behaviour do I need to exhibit to reasonably consider myself a *good* Bayesian? Given that the focus of this volume is Bayesian chronology construction for archaeological applications, we presume the reader to have some knowledge of these approaches and we address our detailed comments primarily to those who use, or are planning to use, these methods to conduct their own analyses. We start, however, by setting the Bayesian paradigm and its application within the wider context of philosophy and science.

2 Philosophy and science

We should explore a few ideas (from Western European traditions of science and philosophy) so that we may continue without amplifying the potential for misunderstanding. Philosophy, literally a 'love of wisdom', is an attempt to fully explore problems within a robust methodology; it is also the historical precursor to the subjects now termed the natural sciences. What became known as the scientific method (the empirical methodology promoted, for example by Francis Bacon in the Novum Organum Scientiarum, 1620) has changed, normatively, over the past hundred years as the belief that an 'objective' scientific practitioner may directly relate observations and generalise them into a 'law', has waned. For a thorough philosophical examination of the 'Received View' and its demise we recommend Suppe (1974).

In the early twentieth century, Einstein was among those to note that "No amount of experimentation can ever prove me right; a single experiment can prove me wrong."(reproduced in Einstein, 2002). He was struggling with the notion that scientific knowledge might be fallible and the distinction between 'induction' and 'deduction' was central to his concern. Deduction, working from general theories to experiments broadly designed to confirm them, had been presumed to be sufficient in scientific endeavour and there was a tradition that theories might inductively arise from the 'clear and distinct ideas' which had categorised 'truth' since the time of Descartes (Third Meditation, in 'Meditations on First Philosophy', 1641). So the scientific method, although differently stated by many scientists, may be briefly characterised as induction leading to hypotheses from which are deduced predictions that are tested by observation. As Box (1976) puts it: "...science is a means whereby learning is achieved, not by mere theoretical speculation on the one hand, nor by the undirected accumulation of practical facts on the other but rather by a motivated *iteration* between theory and practice...".

We are reminded of the way that Arthur Conan Doyle makes his character, Sherlock Holmes, use such disingenuous phrases as: "It is a capital mistake to theorise before one has data. Insensibly one begins to twist facts to suit theories, instead of theories to suit facts."(Conan Doyle, 1892). This intensifies the irony that Holmes almost exclusively uses abductive inference (first referenced in Aristotle's Prior Analytics, republished in Aristotle, 1938), from single instances of data, rather than deduction and ignores the cases where, when quantitative data are in short supply or qualitative data quality is poor (but associated qualitative observations and judgements are well developed) theorising and constructing models before data are obtained may well be the foundations of good, highly principled, subjective reasoning. A practical application of such an approach will be outlined in Section 8.1. For those with no quantitative data, however, abduction has its attractions and it has resurfaced recently as a methodological tool since the social science adoption of qualitative research (see, for example, Reichertz, 2004).

Indeed, at the time that Doyle was writing, it was becoming more tenable that what had been seen as the causative foundations of 'laws' were in fact, as the philosophers Hume (1711–1776) and Kant (1724–1804) had proposed, the product of human psychological effects and the limitations of measurement rather than eternal verities. Frank Ramsey, one of the Cambridge Apostles, continues this investigation in Facts and Propositions (1927) (republished in

Ramsey, 1990). Thus as Poincaré in Science and Hypothesis (1902) asserted, not only was absolute truth not possible within the strictures of science, but 'truths', *per se*, are matters of convenience for scientists rather than being more logically valid than the 'truths' that they replace. The notion of scientific 'laws' as inviolable was coming under intense scrutiny as examples which had stood for 200 years were found wanting. Newton's law of universal gravitation, for example, which as Einstein noted (in his review article of General Relativity in 1916) has inconsistencies due to the deflection of light by the mass of our sun and, as Duhem noted in 1914 (republished in Duhem, 1954) cannot as Newton claimed, be based on Kepler's second and third laws since it contradicts them.

One thing that Einstein had failed to account for, important in the Bayesian framework, is that refuting instances (within a theoretical framework) rather than simply destroying a theoretical under-pinning can, for a practical scientist, lead to improvements of the original hypothesis by refining it. As Einstein was discovering, Newtonian physical theories under-determine the physical universe and uncertainty, rather than being a weakness to be avoided, was becoming acknowledged as central to the scientific pursuit. A sign of this philosophical maturity, away from simplistic mechanistic determinism, is that 'laws' (even where quoted as such for brevity) are now widely thought of as contingent and subject to varying levels of belief that may be best expressed as probabilities subject to statistical reasoning – albeit that their reporting is not often so nuanced.

Alongside this development of philosophical ideas, the necessary probability theory and applied statistical methods for handling and modelling uncertainty were also developing so that, by the late 20th century, several schools of statistical reasoning were well established. Stephen Senn (2011, p. 50) points out that there are four general approaches to statistical reasoning, three of which allow for elements of subjectivity. These approaches contrast Bayesianism (avowedly subjective) with the more classical/frequentist (notionally objective) approaches to statistics but, within each of these, Senn also differentiates those who believe it is possible to interpret derived probabilities as inferences about the 'real' world from those whose focus is solely the current model under investigation. Of these, only those whose focus is on the current model and explore it within the classical statistical framework are operating within the deductive mode and might thus be described as objective. Such analytic *a priori* reasoning, however, leads to truths that are so, solely by virtue of their definition. Newton's first two 'laws' of motion, for example, are inherent in their prior definitions.

In what follows we take an unapologetic, subjective, Bayesian approach to scientific reasoning – not least since we are human, biased and fallible. We do so, in part, because we hold to the relevant philosophical positions outlined here, but also for practical reasons relating to the way in which modern archaeology is practiced (for details see Section 7).

3 Bayesian philosophy of science

Given the large number of existing, good, general accounts of the Bayesian philosophy of science written for a wide range of audiences (e.g. McGrayne, 2011; Lee, 2012; Stone, 2013; von der Linden et al., 2014), here we do not attempt to provide even a summary of the complete paradigm. Instead, we

focus on some key ideas that are particularly pertinent to its principled (or *good*) use in archaeological inference.

Bayes' theorem is often articulated as:

$$Posterior \propto Likelihood \times Prior$$

where *Likelihood* refers to a function (or collection of functions) of the parameters of a statistical model, *Prior* refers to a probabilistic representation of what was known about the parameters before the current data were collected and *Posterior* refers to the result of combining the model, data and the prior knowledge (using statistical inference techniques) to arrive at an updated estimate of the parameter values based on all three. Given these components if we are to make *good* use of the theorem, we must take responsibility for the particular statistical distributions used to construct the likelihood and ensure that our *a priori* knowledge about the parameters of those distributions is appropriately represented. Each of these relies on personal (or collective) judgements and hence subjective assessments of observations or experiences are what turns a merely logically valid conclusion into one in which we can have realistic confidence. Consequently, black-box use of Bayesian methods (other than those for the very simplest problems) is unlikely to be principled.

To take responsibility does not require us to devise or implement models ourselves, but it is necessary that we have a good intuitive feel for what we are aiming to model and the nature of the representation being proposed by those who develop the models and software we plan to use. To achieve this, those providing software have a responsibility to offer natural language (as well as technically correct) explanations of their modelling approach and implementation. While those using or refereeing papers about such software have a responsibility to read the explanations carefully and interact with the author about details that are unclear or seem inappropriate for their needs. The reason this is so important is that the next step involves specifying priors for parameters of the model and, unless we have at least an intuitive understanding of the model and see it as a useful representation of our underlying beliefs, we cannot possibly provide such information reliably.

4 Subjective and objective priors

The choice as to how we specify priors is of necessity a personal one, but one which we must be able to justify and articulate clearly if we are to take full responsibility for the science that we do. There are currently two common approaches, known as subjective and objective Bayes. However, the former is a tautology and the latter an oxymoron. Simply by selecting a Bayesian approach, we have committed to including prior knowledge in the inference process and, in doing so, have selected a subjective scientific framework. To then seek objectivity in prior definition is contradictory.

In practice for many real applications, including archaeological one's, it is far from desirable to claim objectivity since considerable expert knowledge exists and we will do much better science if it can be formalised and included in the data analysis process. Indeed many of us select the Bayesian framework precisely because we seek a means to bring together informative, highly subjective, prior knowledge and more formal, scientific, data in a coherent way. That said, there

are circumstances in which we have little or no informative prior knowledge about key parameters of our model. This occurs most often for what are called *nuisance* parameters i.e. ones that are needed for mathematical or statistical purposes and have no direct counterpart in the reality being modelled. We want the way we model these parameters to have as little impact on the results as possible and so we seek *uninformative* or *vague* priors. In so doing, we are not becoming more objective, we are simply expressing our lack of knowledge in a responsible way. In practice selecting priors with such characteristics can be technically challenging and so it has become a major (pre)occupation of applied Bayesian statisticians – some of whom call themselves objective Bayesians.

From the perspective of a user of Bayesian software, our responsibility is to understand intuitively the nature of the prior distributions that the software developer has coded, to check that for our purposes they seem reasonable and, if not, to consider changing them — perhaps in consultation with the original author/programmer. For vague priors on nuisance parameters, this is often the limit of what is required since software developers typically invest considerable effort in selecting default prior distributions for these parameters to ensure that, under a range of likely data, the priors have negligible impact on the posteriors. However for the parameters that have direct counterparts in reality and about which we are likely to have at least some informative prior knowledge our responsibilities extend further, since for these we will need to select suitable parameterisations of the chosen distributions in order to represent our current state of knowledge. Without such care one invalidates the Bayesian paradigm and so is no longer, as we explore below, strictly engaged in a scientific pursuit.

5 Priors as degrees of belief

There are lots of examples of informative prior knowledge in archaeology since evidence accrues slowly in bits and pieces and it requires expertise to draw it all together. A classic example of this relates to the relative ages for the parameters of a chronological model. Both the model and our prior knowledge of its parameters emerge as we excavate one or more sites and begin to understand the stratigraphic and other chronological relationships. Thus, knowledge is person-specific and time varying. The added complication is that all models are imperfect since they are representations of reality built for a particular purpose. As the statistician George Box said "essentially, all models are wrong, but some are useful" (Box and Draper, 1987, p. 424) and if we use them inappropriately or our needs are different from those of the original developer then we are likely to run into difficulties. Thus we must choose a model that is fit for purpose and choose our prior distributions in a responsible manner.

In selecting prior distributions, we must be clear whether we are seeking to express personal or collective prior knowledge. The distinction is important, since the beliefs held by individuals are typically different from one another and more precise (i.e. less uncertain) than those of a wider group, even one whose members share scientific and other cultural backgrounds. As a result, use of personal prior distributions within a community in which others do not share the same beliefs can lead to accusations of over-fitting (i.e. failure to acknowledge unavoidable sources of uncertainty). Articulating precise personal prior knowledge is, of course, to be encouraged where that knowledge is founded

on clearly articulated expert judgement or other sources of evidence. However, where this is absent, it may well be best to acknowledge the diversity and work closely with those of other opinions to develop a shared, community-wide statement of prior knowledge derived using formal elicitation techniques (see below).

However we derive them, prior distributions are used to represent our personal or collective state of knowledge and so, providing that we explain where our knowledge comes from and why we are representing it the way that we are, we are each entitled to choose different prior distributions. Indeed it would be a surprise if any two archaeologists looking at the same body of knowledge chose to represent it using exactly the same narrative, interpretation, model or prior distributions. For this reason, the key to principled prior definition is the explanation we offer to go alongside the probabilistic choices that we make. Such explanations will be partly generic and relate to the choice between specific probability distributions and for these we may choose to rely in part on the advice of those with more technical, applied statistical, training than ourselves. All explanations of prior selection, however, should contain substantial descriptions that relate to the researcher's real world observations or theories. These should lead to very specific statements about the relative or absolute values of parameters in the form of either clear qualitative statements (e.g. "for the stratigraphic reasons articulated above, the true calendar dates of all of the samples in Phase I must be older than those in Phase II") or direct statistical ones (e.g. "we model our prior knowledge of the length of time elapsed between the stratigraphically older sample A and the younger sample B using a gamma distribution with mean close to zero, but a large variance, to summarise our prior knowledge that whilst the samples are from stratigraphically closely related contexts, there is a small probability that even close stratigraphic relationships on this site indicate long elapsed times").

Such formalism is key because, without it, it is very difficult to avoid double use and hence double weighting of information. For example, since data are often collected before priors are specified, it is difficult to be sure that we have not included information from the data in constructing the prior. The only way to ensure that we do not is to explain, in some detail, how our prior was derived — checking at each step that no recourse to the current data is required.

Thought of in this way, prior distributions represent degrees of belief which are updated using the selected model and the most recently collected data to arrive at posterior distributions which depend on all three. Posterior distributions are then themselves representations of belief which can, in turn, be used as priors for the next piece of research and updated as new information comes to light.

6 Formalising, managing and updating beliefs

For most archaeological purposes knowledge is uncertain because the real world is stochastic (rather than deterministic) and because the information we have access to is a degraded facet of a non-random part of past reality. Given this, we need a way to formalise uncertain knowledge. As outlined above, Bayesians do this using probability distributions and update knowledge using Bayes' theorem.

Unfortunately, probability distributions and in particular conditional prob-

ability distributions (i.e. what we know about one parameter given a value or distribution for another parameter) have properties that are often far from intuitive (e.g. the Monty-Hall problem Selvin, 1975b,a; Rosenhouse, 2009). For this reason encoding one's beliefs using such distributions can be tricky. Indeed, there is a discipline known as 'knowledge elicitation' devoted to helping experts reliably encode their prior beliefs (O'Hagan et al., 2006). Some of these techniques are adopted by those providing software for Bayesian inference in archaeology, but there is still rather too much expected of the users given the acknowledged difficulty of encoding one's own beliefs reliably. This means that some users are adopting default priors without really understanding the choices they have made and others understand the choice, know it is not appropriate, but have few options given that they are not statisticians or coders themselves. More dialogue and work is certainly needed in this area in the coming years.

Even when one has a suitable model and has encoded one's prior knowledge, however, combining probability distributions for all but the simplest models is not trivial. The task is essentially one of integrating over a large number of inter-related and often complex probability distributions and this cannot be done analytically. As a result, until the late 1980s most substantial applied problems were not tractable within the Bayesian framework.

At that time, however, computers became more affordable and computationally intensive simulation techniques began to be used in a wide-range of fields. Statisticians adopted, Monte Carlo (MC) methods that had been widely used in physics and Bayesian statisticians, in particular, adopted a class of these known as Markov-chain Monte Carlo (MCMC) methods (Gelfand and Smith, 1990; Gilks et al., 1996; Robert and Casella, 2010). Both MC and MCMC methods are now widely used in archaeology with the former popular for exploring poorly understood phenomena (e.g. Crema, 2012; Timpson et al., 2014) and the latter for exploring joint probability distributions by sequentially sampling values for individual parameters, conditional on the current values of all of the others.

A host of other simulation-based methods have been explored for Bayesian inference; one of which, Approximate Bayesian Computation (ABC), has proven useful in situations where the likelihood function cannot be evaluated. These methods are mathematically well founded, but since they rely on presumptions and approximations that MCMC methods do not, they add further problems and so should only be used when MCMC is not an option. Such situations are relatively common in biology and so ABC methods are increasingly common in ecology and evolution studies. It is thus no surprise that the first uses in archaeology have been for investigating patterns of evolutionary change (e.g. Crema et al., 2014).

Bayesian chronology construction software, such as OxCal Ramsey (2009) and BCal Buck et al. (1999), adopt MCMC methods and so (given the focus of the current volume) it is these on which we will focus here. These tools are powerful and effective, but bring with them a host of decisions and hence responsibilities for users. There is not space here to discuss all of these in detail. We simply highlight some that all principled users of such methods need to know about and take responsibility for.

MCMC-based methods approximate by repeatedly sampling from the distributions that we are seeking to estimate. Like all sampling methods, we need to be sure that we are sampling from the distribution of interest and have taken a sufficiently large and varied sample to explore it in its entirety. Key concepts

here include the following.

- *Burn-in*: used to refer to samples taken at the start of the simulation process which may not be very representative of the posterior distribution and so are discarded.

- *Convergence*: used to describe the desired state of the sampler i.e. when we are sampling from the correct distribution and so our samples can reliably be used to estimate our posterior.

- *Thinning* and *effective sample size*: inter-related terms used to refer to the process of storing only a sub-set of the samples to leave a sample size that is smaller than the total but nonetheless conveys equivalent information about the underlying distributions. This is important in situations where neighbouring samples in the MCMC simulation are highly correlated and thus contain less information than independent samples of the same size. The goal is to store samples that are as close to independence as possible.

These ideas all require considerable knowledge of MCMC sampling theory and techniques to appreciate in their entirety (for those interested in exploring further we suggest: Gilks et al., 1996; Robert and Casella, 2010), but an intuitive understanding is adequate to make responsible use of well-written software and to adequately document the methods and results obtained. Most software providers, for example, now offer some automated convergence checking which provides guidance about the amount of burn-in and thinning required. Users need simply to understand intuitively the checks being conducted on their behalf, ensure that they are adequate for their own needs and then record the choices made in their writeup so that others may replicate and/or build appropriately on their work.

There is, however, no way automatically to ensure convergence of all parameters of a complex model. As a result, users should undertake reproducibility experiments. This involves making multiple runs of the sampler for each model, with different start values for the sampling chains, to check that the results obtained for key parameters (i.e. the ones that are most important to them and others likely to rely on their work) are the same to an appropriate level of accuracy. *Appropriate* here is of course a relative term. In dating, for example, we will require a different level of accuracy in the historic period from the palaeolithic in order for results to be useful. So our reproducibility experiments, and the accuracy to which we report results, should reflect this.

Since applied Bayesian statistics involves so many personal judgements we all also have responsibility for exploring the sensitivity of the posterior distributions we obtain. We can achieve this by varying our likelihoods and priors and then observing and reporting on the resulting changes in the posteriors. Such investigations are an essential part of any *good* Bayesian analysis since without them we, and others who rely on us, have no idea how robust our results are nor can we safely base any conclusions on them. Should they turn out not to be robust to key decisions we have made, which for large or complex models is not unusual, we will of course have to explore the inductive/deductive hypothesis-observation cycle again to examine why this is the case before deciding on a principled way forward.

7 Bayesian archaeology

Classical (or frequentist) approaches to statistical inference are generally constructed around repeated experiments within a given inductive theoretical framework and rely on phrases like "in the long-run we expect ...". Such arguments might be appropriate in some particular applications in archaeology, for example when considering characteristics of the output from a long-running Medieval pottery production site, but are much harder to justify in general since the 'long-run' here tends to be a substitute term for repeatedly modelling circumstances approaching infinity in one or more dimensions and in prehistory such circumstances are rare. For example, suppose we are seeking to interpret a small collection of cultural artefacts from a single excavation at, say, a short-lived Bronze Age occupation site in the south-west of England. We might think of this as a *one-off experiment* to explore a relatively poorly understood past culture from which little of the total past activity is preserved. In such situations does it ever make sense to say "in the long-run we expect ..."? Our feeling is not, since even if we located and fully excavated all of the remaining well-preserved sites (or even had access to all those ever occupied) our total sample size would still, likely, be very small and the variability between them so great that generalising from one to the totality would be unscientific and very probably also unreasonable.

In addition, in archaeology we typically have very little control over what information we gain access to and when, but we may have alternative theories that compete for a fit to the current data. Much of our information arrives in a haphazard fashion as we do field and laboratory work. To interpret data that arrives in this way within the classical statistical framework, we have a choice. The first option is to interpret each dataset in isolation from the others as they arise and then use a meta-analytic approach to draw together the disparate information at the end. There are well-established protocols for doing this, but considerable information is lost in the process since we combine only the final results from each study, not the raw data. The alternative is to wait until all of the data have been collected and then analyse them together. The difficulty here being that there is no point at which *all* data become available not least since new sites can be discovered or more funding become available. So, we need a way to update knowledge a bit at a time in the light of changing expert opinion and new data — the Bayesian statistical framework is ideally suited to this and the classical one, for structural mathematical reasons rather than ones of ideological preference, is not.

Archaeological science, by its nature, uses a mixture of intellectual approaches: some based in the sciences and some the humanities. We are seeking to blend scientific data with the interpretations, opinions and expertise of human researchers. The Bayesian paradigm offers a framework for formalising this process elegantly and robustly and has been shown, by more than twenty years of illustration, to be a powerful tool in archaeology.

To us these observations, taken together with the fact that Bayesian methods are now relatively easy to learn and to apply, make it surprising that they are applied to so few problems in archaeology. Although there are proof-of-concept examples in several application areas in Buck et al. (1996) and others have also published examples (Orton, 2000; Millard, 2002; Millard and Gowland, 2002; Byers and Roberts, 2003; Millard, 2004, 2005; Finke et al., 2008;

van Leusen et al., 2009; Fernandes et al., 2014), there is really only one application area where Bayesian methods can be said to be routine: absolute, scientific-dating-based chronology construction. There is one other application area with close connections to archaeology, that of phylogeny (both genetic and linguistic), where use of Bayesian methods is also increasingly routine (Drummond et al., 2004; Edwards et al., 2007; Kitchen et al., 2009; Drummond et al., 2012; Bouckaert et al., 2014). Here, however, methodological development was driven largely by the genetics and linguistic research communities. Even areas such as relative dating via seriation and image processing for archaeological field survey data (both considered in Buck et al., 1996) have not, as far as we are aware, seen routine adoption of Bayesian methods.

It seems unlikely that chronologists are the only philosophical Bayesians in the archaeological community, so why are they the only ones routinely using such methods? It may be that they are the only ones with a critical mass of users large enough to invest the time and energy to learn such methods and/or that they are the only ones with obvious, and generally not very contentious, prior information (in the form of stratigraphic information) that they simply cannot afford to ignore. However, there is also a key characteristic of the problem which means that it is obviously a probabilistic one i.e. scientific dates are usually provided to users as probability distributions (in the from of laboratory estimates with standard errors). Given that, in the case of radiocarbon dating, calibration also requires statistical inference, it is then fairly easy to see why some users of Bayesian chronology construction techniques describe themselves as "pragmatic" users of Bayesian methods (e.g. Bayliss and Ramsey, 2004) and why philosophers of science refer to them in much the same way (Steel, 2001).

Some such pragmatic users are also well-informed about the nature of the underlying philosophical and mathematical framework in which they are working and hence do a *good* job. Others are less well-informed and make unsupported assumptions or use only parts of the framework: for example, picking and choosing the tools they find most useful without checking that their theoretical underpinnings are appropriate for the data to hand. In so doing they obtain results that are not robust to statistical scrutiny or are simply not as powerful as they might otherwise have been. Such mis-uses should not deter us from developing and advocating Bayesian methods where we perceive utility, they should simply encourage us to seek to be more careful in their use.

8 Looking to the future

More than twenty years since the first proof-of-concept papers, Bayesian chronology construction is now routine (at least in European prehistory) and, as a result, there is a wealth of literature about its use and application within that particular community. Nonetheless there is potential for much greater benefit, both amongst chronologists and in the wider archaeological community.

8.1 What-if experiments

One powerful, but underused application of the Bayesian framework, is the *what-if* experiment. These were first suggested for chronology construction by Christen and Buck (1998) and have been advocated since, amongst others, by

Bayliss and Ramsey (2004). These involve the simulation of data, prior to or instead of the collection of real data. The benefit is that we can explore the likely gains to be made from data collection at little cost. To utilise this approach one first constructs a model e.g. a relative chronological model based purely on stratigraphic evidence; this can be done before any scientific dating evidence has been obtained. One then uses the model to parameterise data simulation at locations where future absolute dating evidence might be obtained. For example at stratigraphic locations where suitable samples for scientific dating have been obtained.

By then analysing the simulated data in exactly the same way that any future real data would be analysed, we can explore the likely benefit of sending particular groups of samples for dating as opposed to others. In situations where we are keen to achieve a specified accuracy on particular parameters of our model, for example relating to the dates of key events on our site, we can carry out numerous simulation experiments to ascertain which of our datable samples are likely to achieve this and thus target available funds at those samples.

Such *what-if* experiments are cheap, have clear benefits to funding agencies and researchers alike and should help resolve at least some issues without the need for buying real data at all. For example, if the dating accuracy needed cannot be obtained even if we buy scientific dates for all the datable samples from a site, it would surely be better to know this ahead of time and not buy the dates. Given the power of this approach and the fact that software such as OxCal (Ramsey, 2009) has facilitated it for some time, we feel that there is considerable untapped potential here. Certain funding bodies, including the Natural Environment Research Council Radiocarbon Facility already advocate the use of such methods in their FAQ for applicants (Ramsey, 2014) and so presumably they appear in funding applications. Destpite this, although there are published examples, most notably those led by Alex Bayliss (e.g. Bayliss and Ramsey, 2004; Bayliss et al., 2007b; Bayliss and Woodman, 2009), few of these compare the *what-if* experiment directly with the analysis of subsequently obtained real data. More such comparisons would surely be valuable and our feeling is that this approach should become routine.

8.2 Applications beyond chronology

As was clear from Buck et al. (1996), we see real potential for more widespread use of Bayesian methods in archaeology so why have these still to take off? One reason might be that problems, other than chronological and phylogenetic ones, often do not give rise to data which obviously take the form of probability distributions; although many, in fact, do simply because they arise from the stochastic world in which we live. The issue here is that archaeologists rarely have the luxury of repeat observations and so do not readily see their data as they are; samples (albeit often small ones) that could be represented probabilistically. It may simply be that, for greater scientific authenticity, more collaboration is needed with applied statisticians to help devise suitable representations.

Another reason for lack of more widespread adoption might be that, in application areas other than chronology, the associated prior information is often more contentious than that from stratigraphy. However, the facility to explore the consequence of holding different prior opinions is one of the powers of the Bayesian paradigm which is not available within alternative statistical frame-

works. It requires those who hold different opinions to formalise them within the rigour of probability theory and thus to explore explicitly the beliefs they hold. As de Finetti noted in 1931 — probability is required to predict to what extent something is true and since what informs this prediction is a subject with experience and beliefs, "the logical instrument that we need is the subjective theory of probability" (de Finetti, 1989). It also provides the machinery to provide posterior probability statements about the consequences of each person's beliefs on an analysis of the same data and allows us to compare the results quantitatively. In this way we may reach conclusions about the impact of differing opinions on the final interpretations we make. Published examples of such uses are rare, however, and it would be good to see more in the archaeological literature.

8.3 Closing remark

Returning to an issue alluded to in the Abstract, the powerful methods now used so ubiquitously in chronology construction grew from close collaboration between Bayesian statisticians and archaeologists who were intuitively Bayesian, but did not have the formal mathematical training to develop tailored, Bayesian methods for themselves. This close collaboration was the key to developing methods that were not just novel statistically, but were also truly useful to the user community for whom they were designed. Such close collaboration is, however, rare and in danger of being lost altogether unless those who benefit most take action. It is widely acknowledged that there is an international shortage of statisticians (so much so that it is a priority funding area for the UKs Engineering and Physical Sciences Research Council) and so if we want to maintain a connection between the two disciplines we must seek to cultivate it. For archaeological users of statistical tools and methods, this means actively seeking to build strong working relationships with applied statisticians during research design, seeing them as full members of research teams and costing their time into budgets at the funding stage. Without action of this sort we risk losing our place as one of the very few applied scientific communities that are cited by philosophers (e.g. Steel, 2001) as Bayesian.

9 Acknowledgements

Buck is grateful to members of the Natural Environment Research Council-funded BRITICE-CHRONO Team, especially Chris Clark, Richard Chiverrell and Ed Rhodes, for an invitation to and conversations at their November 2014 annual meeting at the Palace Hotel in Buxton, UK. The authors are also grateful to two World Archaeology reviewers for their comments on an earlier draft of the paper which led to important revisions.

References

Aristotle (1938). Prior Analytics (Hugh Tredennick [trans.]). In *Aristotle, Volume 1*, pages 181–531. Loeb Classical Library, William Heinemann, London, UK.

Bayliss, A. (2009). Rolling out revolution: using radiocarbon dating in archaeology. *Radiocarbon*, 51(1):123–47.

Bayliss, A., Bronk Ramsey, C., Van der Plicht, J., and Whittle, A. (2007a). Bradshaw and bayes: towards a timetable for the neolithic. *Cambridge Archaeological Journal*, 17(S1):1–28.

Bayliss, A. and Ramsey, C. B. (2004). Pragmatic Bayesians: a decade of integrating radiocarbon dates into chronological models. In Buck, C. and A.R., M., editors, *Tools for Constructing Chronologies*, pages 25–41. Springer.

Bayliss, A., Ramsey, C. B., van der Plicht, J., and Whittle, A. (2007b). Bradshaw and Bayes: towards a timetable for the Neolithic. *Cambridge Archaeological Journal*, 17:1–28.

Bayliss, A. and Woodman, P. (2009). A new Bayesian chronology for Mesolithic occupation at Mount Sandel, Northern Ireland. *Proceedings of the Prehistoric Society*, 75:101–123.

Beaumont, M. A. and Rannala, B. (2004). The Bayesian revolution in genetics. *Nature Reviews Genetics*, 5:251–261.

Bernardo, J. M. and Smith, A. F. M. (1994). *Bayesian Theory*. John Wiley & Sons, New York.

Blaauw, M. and Christen, J. A. (2011). Flexible paleoclimate age-depth models using an autoregressive gamma process. *Bayesian Analysis*, 6(3):457–474.

Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., Wu, C.-H., Xie, D., Suchard, M., Rambaut, A., and Drummond, A. J. (2014). BEAST 2: A Software Platform for Bayesian Evolutionary Analysis. *PLoS Computational Biology*, 10(4). e1003537. doi:10.1371/journal.pcbi.1003537.

Box, G. E. P. (1976). Science and statistics. *Journal of the American Statistical Association*, 71(356):791–799.

Box, G. E. P. and Draper, N. R., editors (1987). *Empirical Model-Building and Response Surfaces*. John Wiley & Sons, New York.

Brooks, S. P. (2003). Bayesian computation: a statistical revolution. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 361(1813):2681–2697.

Buck, C. E., Cavanagh, W. G., and Litton, C. D. (1996). *The Bayesian Approach to Interpreting Archaeological Data*. Wiley, Chichester.

Buck, C. E., Christen, J. A., and James, G. N. (1999). BCal: an on-line Bayesian radiocarbon calibration tool. *Internet Archaeology*, 7. `http://intarch.ac.uk/journal/issue7/buck/`.

Buck, C. E., Litton, C. D., and Smith, A. F. (1992). Calibration of radiocarbon results pertaining to related archaeological events. *Journal of Archaeological Science*, 19(5):497–512.

Byers, S. N. and Roberts, C. A. (2003). Bayes' theorem in paleopathological diagnosis. *American Journal of Physical Anthropology*, 121:1–9.

Christen, J. A. and Buck, C. E. (1998). Sample selection in radiocarbon dating. *Applied Statistics*, 47:543–557.

Conan Doyle, A. (1892). A Scandal in Bohemia. In *The Adventures of Sherlock Holmes*. George Newnes, London.

Crema, E., Edinborough, K., Kerig, T., and Shennan, S. (2014). An Approximate Bayesian Computation approach for inferring patterns of cultural evolutionary change. *Journal of Archaeological Science*, 50:160–170.

Crema, E. R. (2012). Modelling temporal uncertainty in archaeological analysis. *Journal of Archaeological Method and Theory*, 19:440–461.

de Finetti, B. (1989). Probabilism, *Logos*, **14**, pp 163–219, 1931, English translation. *Erkenntnis*, 31:169–223.

Drummond, A., Nicholls, G. K., Rodrigo, A. G., and Solomon, W. (2004). Genealogies from time-stamped sequence data. In Buck, C. E. and Millard, A. R., editors, *Tools for Constructing Chronologies: crossing disciplinary boundaries*, pages 149–171. Springer-Verlag, London.

Drummond, A., Suchard, M., Xie, D., and Rambaut, A. (2012). Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Molecular Biology and Evolution*, 29(8):1969–1973.

Duhem, P. (1954). *La Théorie Physique: son Objet et sa Structure*. Princeton University Press, Princeton, 2nd edition. Translation by Philip P. Wiener.

Edwards, C. J., Bollongino, R., Scheu, A., Chamberlain, A., Tresset, A., Vigne, J.-D., Baird, J. F., Larson, G., Ho, S. Y., Heupink, T. H., et al. (2007). Mitochondrial DNA analysis shows a Near Eastern Neolithic origin for domestic cattle and no indication of domestication of European aurochs. *Proceedings of the Royal Society B: Biological Sciences*, 274(1616):1377–1385.

Einstein, A. (2002). Induction and deduction in physics. In Janssen, M., Schulmann, R., Illy, J., Lehner, C., and Buchwald, D. K., editors, *The Collected Papers of Albert Einstein, Volume 7: The Berlin Years: Writings, 1918-1921. (English translation of selected texts)*, pages 218–236. Princeton University Press.

Fernandes, R., Millard, A. R., Brabec, M., Nadeau, M.-J., and Grootes, P. (2014). Food Reconstruction Using Isotopic Transferred Signals (FRUITS): A Bayesian Model for Diet Reconstruction. *PLoS ONE*, 9.

Finke, P. A., Meylemans, E., and Van de Wauw, J. (2008). Mapping the possible occurrence of archaeological sites by Bayesian inference. *Journal of Archaeological Science*, 35:2786–2796.

Finkelstein, I. and Piasetzky, E. (2010). Radiocarbon dating the Iron Age in the Levant: a Bayesian model for six ceramic phases and six transitions. *Antiquity*, 84(324):374–85.

Gelfand, A. E. and Smith, A. F. M. (1990). Sampling based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85:398–409.

Gilks, W., Richardson, S., and Spiegelhalter, D., editors (1996). *Markov Chain Monte Carlo in Practice*. Chapman and Hall, London.

Haraway, D. (1988). Situated knowledges: the science question in feminism and the privilege of partial perspective. *Feminist Studies*, 14(3):575–599.

Haslett, J. and Parnell, A. C. (2008). A simple monotone process with application to radiocarbon-dated depth chronologies. *Journal of the Royal Statistical Society, Series C*, 57:399–418.

Howson, C. and Urbach, P. (1993). *Scientific Reasoning: the Bayesian approach*. Open Court, Illinois, second edition.

Kaplan, D. (2014). *Bayesian statistics for the social sciences*. Guilford Publications.

Kéry, M. and Schaub, M. (2012). *Bayesian population analysis using WinBUGS: a hierarchical perspective*. Academic Press.

Kitchen, A., Ehret, C., Assefa, S., and Mulligan, C. J. (2009). Bayesian phylogenetic analysis of Semitic languages identifies an Early Bronze Age origin of Semitic in the Near East. *Proceedings of the Royal Society B: Biological Sciences*.

Koop, G., Poirier, D. J., and Tobias, J. L. (2007). *Bayesian econometric methods*, volume 7. Cambridge University Press.

Kruschke, J. K., Aguinis, H., and Joo, H. (2012). The time has come Bayesian methods for data analysis in the organizational sciences. *Organizational Research Methods*, 15(4):722–752.

Lee, P. M. (2012). *Bayesian statistics: an introduction*. John Wiley & Sons.

Lunn, D., Spiegelhalter, D., Thomas, A., and Best, N. (2009). The BUGS project: evolution, critique and future directions (with discussion). *Statistics in Medicine*, 28:3049–3067.

Lunn, D. J., Thomas, A., Best, N., , and Spiegelhalter, D. (2000). WinBUGS – a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing*, 10:325–337.

Manning, S. W. (2007). Preface. Beyond Dates to Chronology: Rethinking the Neolithic-Chalcolithic Levant. *Paléorient*, 33(1):5–10.

Mayo, D. G. and Spanos, A., editors (2010). *Error and Inference: Recent Exchanges on Experimental Reasoning, Reliability, and the Objectivity and Rationality of Science*. Cambridge University Press, Cambridge.

McGrayne, S. B. (2011). *The theory that would not die: how Bayes' rule cracked the enigma code, hunted down Russian submarines, & emerged triumphant from two centuries of controversy*. Yale University Press.

Millard, A. (2002). Bayesian approach to sapwood estimates and felling dates in dendrochronology. *Archaeometry*, 44(1):137–143.

Millard, A. and Gowland, R. (2002). A Bayesian approach to the estimation of the age of humans from tooth development and wear. *Archeologia e Calcolatori*, (XIII):197–210.

Millard, A. R. (2004). Taking Bayes beyond radiocarbon: Bayesian approaches to some other chronometric methods. In Buck, C. and A.R., M., editors, *Tools for Constructing Chronologies*, pages 231–248. Springer.

Millard, A. R. (2005). What can bayesian statistics do for archaeological predictive modelling? In van Leusen, M. and Kamermans, H., editors, *Predictive modelling for archaeological heritage management: a research agenda*, pages 169–182. Rijkdienst voor het Oudheidkundig Bodemonderzoek, Amersfoort.

O'Hagan, A., Buck, C. E., Daneshkhah, A., Eiser, R., Garthwaite, P., Jenkinson, D., Oakley, J., and Rakow, T. (2006). *Uncertain Judgements: Eliciting Expert Probabilities*. Wiley, London.

Orton, C. (2000). A bayesian approach to a problem of archaeological site evaluation. In Lockyear, K., Sly, T., and Mihailescu-Bîrliba, V., editors, *Computer Applications and Quantitative Methods in Archaeology CAA 96*, pages 1–7. Edîtura Demiurg and Archaeopress, Iasi and Oxford.

Ramsey, C. B. (2009). Bayesian analysis of radiocarbon dates. *Radiocarbon*, 51(1):337–360.

Ramsey, C. B. (2014). Frequently Asked Questions: How many dates do I need for my site? `http://www.c14.org.uk/embed.php?File=nercfaq.html#How_many_dates_do_I_need_for_my_site`. NERC Radiocarbon Facility, accessed 21 November 2014.

Ramsey, F. P. (1990). Facts and propositions, aristotelian society supplementary volume. In Mellor, D. H., editor, *Philosophical Papers*, volume 7, pages 34–51. Cambridge University Press, Cambridge.

Reichertz, J. (2004). Abduction, deduction and induction in qualitative research. In Flick, U., von Kardoff, E., and Steinke, I., editors, *A Companion to Qualitative Research*, pages 159–164. Sage Publications.

Robert, C. and Casella, G. (2010). *Introducing Monte Carlo Methods with R*. Springer-Verlag, New York.

Rosenhouse, J. (2009). *The Monty Hall Problem: The Remarkable Story of Math's Most Contentious Brain Teaser*. Oxford University Press.

Selvin, S. (1975a). On the Monty Hall problem. *American Statistician*, 29(3):134.

Selvin, S. (1975b). A problem in probability. *American Statistician*, 29(1):67.

Senn, S. (2011). You may believe you are a Bayesian but you are probably wrong. In Mayo, D. G., Spanos, A., and Staley, K. W., editors, *Rationality, Markets and Morals, Special Topic: Statistical Science and Philosophy of Science*, volume 2, pages 48–66.

Shultz, T. R. (2007). The Bayesian revolution approaches psychological development. *Developmental science*, 10(3):357–364.

Smyth, J. (2013). Tides of change? The house through the Irish Neolithic. In *Tracking the Neolithic House in Europe*, pages 301–327. Springer.

Steel, D. (2001). Bayesian statistics in radiocarbon calibration. *Philosophy of Science*, 68:S153–S164. Copied.

Stone, J. V. (2013). *Bayes' rule: a tutorial introduction to Bayesian analysis*. JV Stone.

Suppe, F., editor (1974). *The Structure of Scientific Theories*. University of Illinois Press.

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

van Leusen, M., Millard, A., and Ducke, B. (2009). Dealing with uncertainty in archaeological predictive modelling. In Kamermans, H., van Leusen, M., and Verhagen, P., editors, *Predictions and Risk Management: Alternatives to Current Archaeological Practice*, pages 123–160. Leiden University Press, Leiden.

von der Linden, W., Dose, V., and von Toussaint, U. (2014). *Bayesian Probability Theory: Applications in the Physical Sciences*. Cambridge University Press.

Whittle, A. and Bayliss, A. (2007). The times of their lives: from chronological precision to kinds of history and change. *Cambridge Archaeological Journal*, 17:21–28.

Whittle, A. W. R., Healy, F. M. A., and Bayliss, A. (2011). *Gathering time: dating the early Neolithic enclosures of southern Britain and Ireland*. Oxbow Books.

Wicks, K., Pirie, A., and Mithen, S. (2014). Settlement patterns in the late mesolithic of western scotland: the implications of bayesian analysis of radiocarbon dates and inter-site technological comparisons. *Journal of Archaeological Science*, 41:406–422.

Wilkinson, D. J. (2007). Bayesian methods in bioinformatics and computational systems biology. *Briefings in bioinformatics*, 8(2):109–116.