Contents lists available at ScienceDirect

Journal of Archaeological Science

journal homepage: www.elsevier.com/locate/jas

Statistical modelling in archaeology: some recent trends and future perspectives

Enrico R. Crema [a,b]

[a] Department of Archaeology, University of Cambridge, CB2 3DZ, UK
[b] McDonald Institute for Archaeological Research, University of Cambridge, CB2 3ER, UK

ARTICLE INFO

Keywords:
Model thinking
Statistical inference
Generative inference
Measurement error
Multilevel models
Uncertainty

ABSTRACT

This paper reviews the application of statistical models in archaeology in the last decade, focusing in particular on multilevel models, statistical treatment of missing data and measurement error, and simulation-based generative inference. These techniques are designed to 1) account for the nested and hierarchical nature of the archaeological record, 2) formally integrate different forms of data uncertainties, and 3) provide a more direct inferential link between formal theory and observational data. The extent to which archaeology has engaged with these methods is variable, but it can be argued that none are currently regarded as part of the standard analytical toolkit in quantitative archaeology. The objective of this paper is to promote awareness of the existence of these techniques and highlight the consequences of ignoring the underlying problems that these statistical methods can address.

1. Introduction

Statistical modelling and, more broadly, quantitative methods have always occupied an idiosyncratic niche in the discipline. The use of statistical techniques in archaeology goes back to the middle of the 20th century (Bordes, 1953; Brainerd, 1951; Spaulding, 1953), and since then, countless edited volumes (Barcelo and Bogdanovic, 2015; Heizer and Cooke, 1960; Leonard and Jones, 1989), monographs (Buck et al., 1996; Orton, 2012), and manuals (Carlson, 2017; Drennan, 2009; Fletcher and Lock, 2005; Shennan, 1997; Van Pool and Leonard, 2011), as well as review articles reflecting its role in the wider disciplines (Aldenderfer, 1998; Ammerman, 1992; Buck and Meson, 2015; Cowgill, 2015; Otárola-Castillo and Torquato, 2018; Thomas, 1978), have been published. Yet, quantitative methods are still far from being unequivocally considered an integral part of our discipline. Statistical training in undergraduate and graduate programs is still limited and treated as an optional module offered at the Master's level (see Vaiglova, 2025 for detailed discussion on the importance of statistical training in archaeological science); often perceived to be relevant only if one wants to venture into an archaeological science curriculum and almost deemed unnecessary for those pursuing other aspects of the field. It comes as no surprise that compared to adjacent disciplines such as biological anthropology, geography, or ecology, the average level of statistical sophistication offered by archaeological research is below the bar.

Still, the application of quantitative methods in archaeology continues to grow, perhaps faster than ever. This upward trajectory most directly results from a number of concurrent factors. The so-called 'Third Science Revolution' (Kristiansen, 2021), the open science movement (Marwick et al., 2017), and the increased availability of legacy data (Bevan, 2015) prompting synthetic (Altschul et al., 2018) and comparative (Drennan et al., 2011) archaeology are perhaps the most prominent drivers behind the tangible increase of computational and quantitative methods in archaeology. The open science movement, in particular, has promoted the practice of sharing computer code and raw data along with published papers, allowing practioners to learn and apply more complex methods more rapidly than ever. Despite this positive trend, some of the warnings and precautions advocated by early quantitative archaeologists are still valid. Many of the issues highlighted by David Hurst Thomas nearly half a century ago (Thomas, 1978) are painfully still relevant, and as ever, the democratisation of sophisticated methods goes in tandem with an increase in the misuse and abuse of inferential techniques (Smith and Sandbrink, 2022).

This article is part of a special issue entitled: <Special issue full title (fetch from PTS); guest edited by editor's name (fetch from PTS)> published in <journal name>. Example: This article is part of a special issue entitled: Next Generation Archaeological Science: Papers in Honour of Robin Torrence; guest edited by Efthymia Nikita and Marcos Martinón-Torres.

E-mail address: erc62@cam.ac.uk.

https://doi.org/10.1016/j.jas.2025.106295
Received 5 February 2025; Received in revised form 20 May 2025; Accepted 12 June 2025

0305-4403/© 2025 Published by Elsevier Ltd.

As noted above, critical reflections on the role of quantitative methods in archaeology have occurred many times in the past and certainly will not cease in the future. This paper follows suit with such a recurrent practice of self-reflection, but placing a particular emphasis on *statistical modelling*, reviewing three promising techniques that address some long-standing challenges in quantitative archaeology. Section 2 will highlight the benefits of *multilevel models* and how they can offer an explicit approach for tackling the hierarchichally nested nature of the archaeological record but also to model the heteoregeneity in human behaviour that we ultimately wish to describe; section 3 will cover statistical models designed to properly handle *missing data and measurement error*, but also the importance of measuring, communicating, and intgerating uncertainty in our analysis; and finally section 4 will overview the possibilities and the challenges offered by *simulation-based inference* and its potential to bridge theory and observational studies.

Before I focus on the themes detailed above, it is necessary to briefly define what is meant by statistical modelling in this context. In a nutshell, a statistical model can be described as “a set of probability distributions on the sample space **S**” (McCullagh, 2002, p. 1225); in other words, a probabilistic description of all possible outcomes (observed and non) of a particular system of interest. A central element here is the use of *probability distributions*, mathematical models describing the probabilities of events and observations given a set of *parameters*. These parameters are the primary focus of *statistical inference*, whether we wish to infer precise and accurate estimates of when a particular technology was introduced, determine the extent by which specific environmental factors led to a higher concentration of human occupation in a given region, or quantify the nature of the association between diet and social status inferred from burial practices. While this principle applies to any inferential technique, from a simple Chi-square test to sophisticated Generalized additive mixed effect models, I will focus here primarily applications where the core objective is to describe observed variation in the archaeological record as a function of model parameters and covariates (aka independent variables or predictors).

2. Multilevel modelling

Any standard introductory textbook on statistics warns the uninitiated that proper inference requires samples to be *random* and *independent*. Indeed, designing appropriate sampling strategies to satisfy these assumptions and, more broadly, to offer robust ground for statistical inference is an integral part of any modern science, and we are not certainly short of subject-specific treatise in archaeology (Banning, 2021; Comer et al., 2023; Orton, 2012; Wells, 2010). However, observational studies in archaeology are often, and perhaps even increasingly, opportunistic — samples collected for entirely different purposes are routinely aggregated to pursue work of synthesis with increasingly broader temporal and spatial scope. While these efforts can bring new life to legacy datasets and offer unique opportunities to engage with the ‘big’ questions of our discipline (Kintigh et al., 2014), they are also subject to new inferential challenges. Because original sampling strategies were designed to tackle specific objectives, regional and cross-regional datasets are characterised by different retrieval methods (e.g. floatation vs dry sieving), measurement protocols (e.g. absolute vs relative chronology), and sampling intensity. These factors potentially contribute, sometimes in a substantial way, to the variability we observe in the archaeological data. Samples coming from specific sites might share unmodelled characteristics that can bias our interpretation when we analyse data from different sites in a regional studies naively ignoring the clustered nature of our record. Regional and cross-regional studies are also affected by spatial and phylogenetic autocorrelation (aka Galton’s problem) — samples close in space (e.g. same site, same region) are statistically non-independent and ignoring this factor can lead to biased conclusions. For example, estimating the origination date of a particular phenomenon (e.g. the appearance of a domesticated crop) often entails statistical examination of the earliest dated samples using methods such as Bayesian phase models (e.g. Leipe et al., 2019) or optimal linear estimation (Key et al., 2021). These techniques can be extremely powerful but can potentially provide biased estimates in the presence of strong sample imbalance; determining the origination date from 20 radiocarbon dates from 20 different archaeological sites will offer a less biased estimate compared to the same number of samples collected from the same archaeological layer in a single site. Aggregate data can also be subject to inferential biases derived by the so-called *Simpson’s paradox*, where patterns observed at a sub-group level do not manifest, or even show opposite signature when groups are combined (see Table 1).

Multilevel models (aka *hierarchical* or *mixed-effect* models) offer a well-established suite of statistical techniques that are designed to address this issue. In a nutshel, these models can be considered a generalisation of linear regression where model parameters (e.g. intercept, slope) vary between observations based on different levels of data structure. Thus, for example, one could model the association between wealth or status (inferred from burial goods) and diet (inferred from stable isotope analysis; e.g. Privat et al., 2002) whilst accounting for which sites each individual are from. In a standard regression model, the association between the two variables would be captured by two parameters (an intercept and slope, Fig. 1-a), whilst *multilevel* models can portray this relationship as a distribution of parameters, whereby each archaeological site will have its unique combination of intercept and slope (Fig. 1-b). The model can thus simultaneously capture the general relationship between the two variables, as well how such association varies across different sites. A systematic introduction of multilevel models is beyond the scope of this paper, but interested readers can consult McElreath’s manual on Bayesian statistics (McElreath, 2020) and the archaeology-focused introduction by Fernée and Trimmis (2021).

Using multilevel models offers several advantages over standard linear regression, so much so that some authors (e.g. McElreath, 2020) argue that this should be the default approach for statistical modelling. For example, the ability to formally account for natural clusters in the data provides a straightforward solution to sample imbalance and can benefit from *partial pooling*. Conventional approaches to the existence of groups or natural clusters in a dataset typically involve: 1) ignoring such structure (i.e. *complete pooling*; Fig. 1-a), potentially disregarding differences between groups and increasing biases introduced by sample imbalance (e.g. a group with a large sample and with an unusual association between dependent and independent variables could bias the overall estimate, see also Table 1 above), or 2) treating the group or cluster as a covariate (i.e. *no pooling*), thus providing group-specific estimates but disregarding information from the general trends and without the possibility to make predictions for a hypothetical

Table 1
An illustration of Simpson’s paradox on a hypothetical study looking at proportion of burial with or without grave goods. When the dataset is examined in its aggregate form, the percentage of male individuals with grave goods is higher than female individuals (67.84 % vs 44.97 %, $\chi^2 = 98.847$, d.f. = 1, p-value <0.0001). However at the site level the proportion of female individuals with grave goods is always higher than male individuals. The paradox arises from sample imbalance, with sites with higher percentage of female burial with burial goods also being the ones with a smaller number of female individuals (i.e. sites A,D, and E).

	Male			Female		
	n	with grave goods	%	n	with grave goods	%
Site A	723	500	69.16	70	64	91.43
Site B	87	18	20.69	345	91	26.38
Site C	65	23	35.38	178	67	37.64
Site D	331	311	84.59	80	74	92.5
Site E	134	88	65.7	23	17	73.91
Total	1340	909	67.84	696	313	44.97

unobserved group (Fig. 1-d). Partial pooling offers an intermediate solution, where the variability between groups is directly modelled (Fig. 1-b, Fig. 1-c) with inferences on particular groups informed by the rest of the samples. To put in simpler terms, multilevel models provide both an approach to provide a better understanding of the variability at higher scales (e.g. how does the relationship between diet and wealth varies across sites), but also model such variability to improve the inference at a lower scale (e.g. a better estimate on the relationship between diet and wealth at a specific site).

The ability to directly model variability offers other advantages, making multilevel models a powerful and flexible inferential tool. Aside from being able to model complex hierarchical structures in the data, multilevel models can also be used to model instances where the same measurement is taken from the same individual multiple times (i.e. *repeated measure* models), account for measurement error (e.g. Error-in-variable models., section 3.2 below), or even infer how variations between groups are conditioned by factors such as geographic or phylogenetic distance (e.g. Gaussian Process models). While these additional layers of complexity can lead to computational and interpretative challenges, the benefits offered by its ability to model different sources of variation explicitly make multilevel models an essential tool of modern statistics.

Applications of multilevel models in archaeology are still few compared to other disciplines where it has become a standard toolkit. Nonetheless, the last few years saw a growing number of applications across the archaeological sciences, enabling, for example, to model compositional variability of goldwork in pre-Hispanic Colombia (Vieri et al., 2025) or examine variation in the relative proportion of ovicaprids in southern Italy during the 1st millennium BC (Ragno, 2024). Other examples can be found in stable isotope analysis (Perri et al., 2019), osteoarchaeology (Alonso-Llamazares et al., 2022; Rosenstock et al., 2019; Wallace et al., 2020), age-depth models (Heegaard et al., 2005), cultural chronology (Banks et al., 2019), diffusion analysis (Crema et al., 2022, 2024), ethnoarchaeology (Bevan et al., 2024), palaeoeconomy (Kohler et al., 2025), zooarchaeology (Wolfhagen, 2024), and palaeodemography (Riris et al., 2024). The range of applications is a testament to how multilevel models offer a solution to problems that are present across different archaeological areas of research. Notably, many of these studies examine legacy datasets covering a large geographic and chronological span, showcasing how multi-level models are particularly well suited in these contexts.

3. Missing data, measurement error, and uncertainty

As a discipline that seeks to recover human behaviour patterns from 'indirect traces in bad samples' (Clarke, 1973, p. 17), archaeologists are

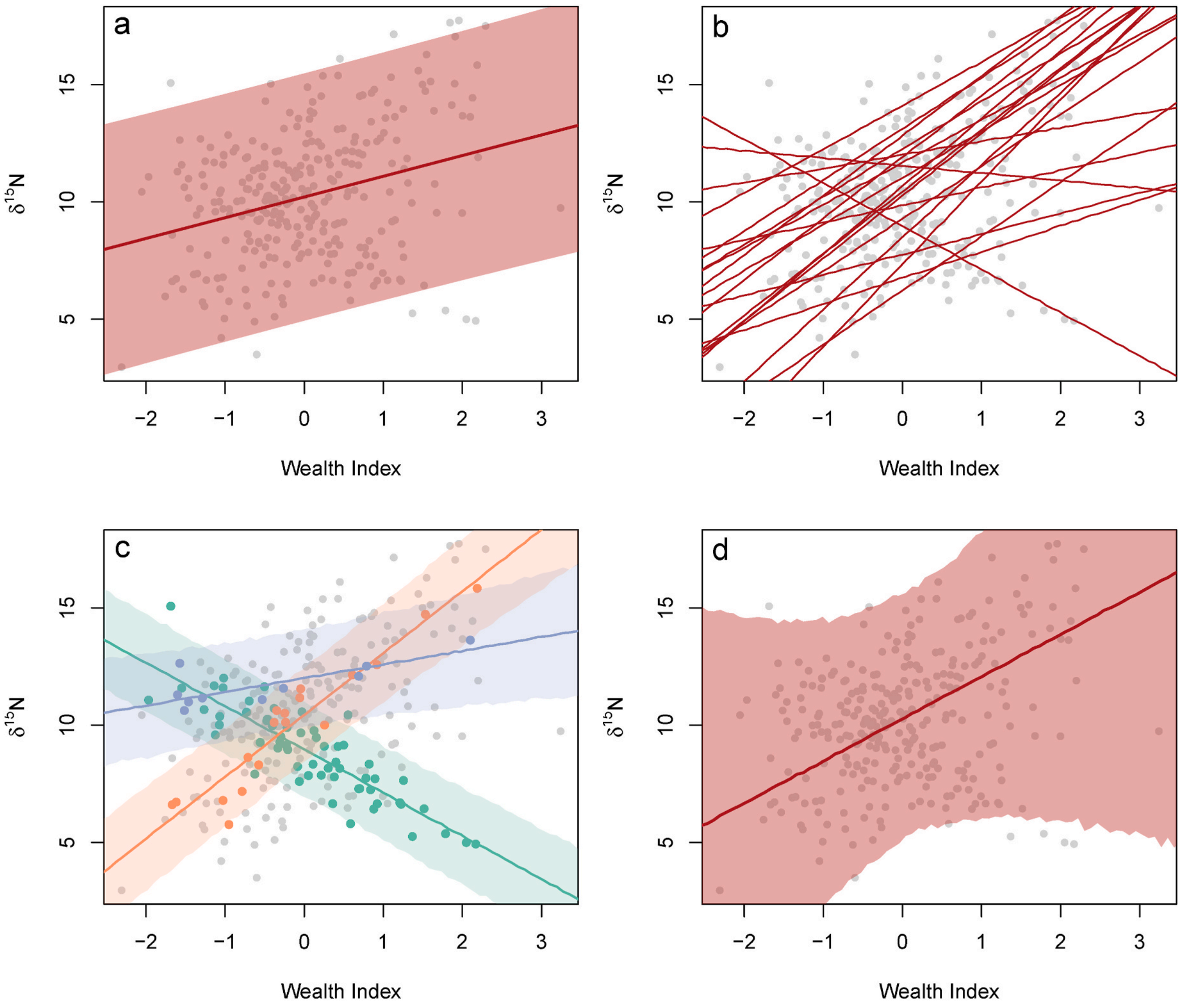

Fig. 1. Analysis of a simulated dataset portraying a hypothetical relationship between a standardised proxy of wealth and diet (nitrogen isotope ratio) for a sample of 294 individuals from 20 different archaeological sites: (a) fitted model and 95 % prediction interval according to a complete pooling standard linear regression model ($\delta^{15}N \sim 1+wealth$); (b) fitted model for each site according to a multilevel model with random slope and intercept ($\delta^{15}N \sim 1+wealth + (1+wealth|site)$); (c) comparison of the prediction interval for three out of the twenty archaeological sites according to the same multilevel model, showing in one case a negative relationship between the two variables; (d) prediction interval for samples from a hypothetical new site according to the same multilevel model.

well-acquainted with the problem of small sample sizes, missing data, and different forms of uncertainties associated with what we directly or indirectly intend to measure. Yet, while philosophical reflections on the role of uncertainty in the discipline have been made from different theoretical perspectives (Gero, 2007; Gonzalez-Perez et al., 2023; Sørensen, 2016; van der Leeuw, 2016), the development and applications of statistical approaches designed to account for these limitations have been somewhat uneven, with some areas of applications ahead of others despite fundamentally (at least in mathematical terms) sharing the same problem. Calls for increased awareness in handling sources of uncertainty, such as measurement error and missing data, have nonetheless been raised in different research areas, from stable isotope research (Jardine and Cunjak, 2005) to skyscape archaeology (Silva, 2019, 2020). Still, universal reflections and practical recommendations are lacking.

Part of the reason why we lack a coherent and systematic approach stems from the fact that uncertainty is present in different forms at different stages of archaeological practice, from data retrieval and measurement to the interpretation and dissemination of analytical output. An ideal pipeline requires the inheritance of the quantified uncertainties at each stage, enabling a formal link between issues of missing data or measurement errors and our parameter estimates and conclusions. In practice, however, uncertainty is often perceived as a limitation rather than the extent of our knowledge, and as such, even if this is appropriately measured, it is not necessarily accounted for in the next stage of the inferential pipeline. Negligence, malpractices, or even a lack of interest in properly accounting for uncertainty stems perhaps from the tacit assumption that the consequences in ignoring these aspects are minimal. Yet, ignoring issues such as measurement error does not just lead to a reduction in precision (i.e. wider confidence interval in parameter estimates) but also in accuracy (i.e. incorrect inference), potentially leading to incorrect conclusions.

3.1. Missing data

Missing data represents a common challenge in archaeological inference, where we often have to face the consequences of a variety of depositional and post-depositional processes that lead to partial or even complete loss of information. Indeed, archaeologists are accustomed to routinely make *inferences from absence*, an epistemic standpoint that is considered a form of logical fallacy in most scientific disciplines, which nonetheless is unavoidable and sometimes even justifiable in our field (Wallach, 2019 for an extensive discussion on this point; see also below). However, the problem of missing data is certainly not unique to our field, and scholars across a wide range of disciplines commonly face the challenge of datasets where values for one or more variables in an observation are simply unavailable. The most conventional solution for handling these instances is to remove any observations with missing data from the sample. This approach, known as *listwise deletion,* is so common that many statistical software applications adopt them by default (e.g. the *lm*() function in R), and researchers often fail to justify, or worse, even mention, that such a decision has taken place. At best, when the processes leading to missing data are completely random (see below), listwise deletion can decrease the power of statistical analyses, increasing the chance of making false negative statements. However, depending on how and why data are missing, listwise deletion can lead to biased estimates (Allison, 2002).

The alternative to removing missing data involves some form of *imputation*, where missing values are replaced with some reasonable guess, and analyses are carried out as if the data were complete. To some extent, *inference from absence* is a special form of imputation where missing data are effectively replaced with zeroes. Wallach (2019) argues that this approach may be justified in archaeological contexts under the premise that (1) human presence generally leaves a strong footprint and that (2) many types of material remains have a high degree of survivability. While both conditions are met in some circumstances (e.g. the complete absence of major urban sites in a region), interpreting *absence of evidence* as *evidence of absence* can also have devastating consequences. For example, a recent paper by Whitehouse et al. (2019) analysed archaeological evidence of so-called 'moralising gods' to investigate whether the presence of supernatural agents punishing free riders followed or promoted the emergence of complex societies. Their conclusions supported the former, but closer inspection of their code showed that missing data on moralising gods were interpreted as evidence of an absence of supernatural punishment. The problem, however, is that evidence supporting the presence (or absence) of moralising god requires written sources, and as such early societies are more likely to have missing data. Subsequent reanalysis of the same dataset (Beheim et al., 2021) employing various statistical approaches for treating missing data led to different results, in some cases providing support for conclusions that were opposite to what was claimed in the original paper. Similar analyses focused on longer chronological time-span are particularly prone to instances where the probability $P_{missing}$ of a value in the data is missing is conditional to time itself. A typical example is the analyses of time-frequency data (e.g. the use of ^{14}C dates to estimate past demographic fluctuations), where time-dependent destructive processes can potentially lead to biased inference (Surovell and Brantingham, 2007).

The examples above hint at the relevance of understanding what conditions $P_{missing}$. A common classification scheme employed by statisticians distinguishes between *Missing completely at random* (MCAR), *Missing at random* (MAR), and *Missing not at random* (MNAR), depending on the relationship between $P_{missing}$ and the variables of interest (Rubin, 1976). MCAR describes instances where $P_{missing}$ is independent of any of the variables of interest, in which case listwise deletion is acceptable, and inference from absence can be cautiously supported when Wallach's two assumptions are met. MAR refers instead to instances where $P_{missing}$ does not depend on the unobserved values but does depend on observed ones. For example, when examining ceramic decorations, information from fragile, thin-walled vessels might not be available due to higher levels of fragmentation. In this case, the $P_{missing}$ of decorative traits depends on a variable that can be observed (the thickness of the vessel). Lastly, under MNAR, $P_{missing}$ is conditional to both observed and non-observed variables. MNAR are particularly problematic as the source of the missing values is not observed (e.g. consider an investigation on wealth inequality based on burial goods where information from the wealthiest tombs is unavailable because of looting). Statistical solutions like imputation methods rely on the researcher being able to classify their particular research context to one of these three levels.

As noted above, the most common treatment for missing data is imputation. Archaeological applications of imputation methods are becoming more frequent and are even featured in some generalist manuals (Baxter, 2003; Carlson, 2017). Yet, a recent systematic survey of over 950 bioarchaeology articles published between 2011 and 2020 (Wissler et al., 2022a) shows that less than a third engaged in some way with missing data, and only 43 papers performed some form of imputation. While a survey on the broader field of archaeological science is not available, statistical treatment of missing data is likely even less common in other areas of applications. The problem is further exacerbated by the fact that there is no single statistical approach to the missing data problem. Indeed, all review articles comparing alternative algorithms (Pang and Liu, 2023; Ryan-Despraz and Wissler, 2024; Wissler et al., 2022b) effectively acknowledge that a single solution does not exist, and researchers should account for the nature of data investigated, the type of missingness, the strength and the weakness of different imputation methods, and the research question posed.

It is worth noting, however, that despite the lack of a discipline-wise awareness of the importance of proper management and treatment of missing data, many of the solutions developed in other disciplines are directly applicable to archaeology and have been implemented in areas outside bioarchaeology (Fanta et al., 2020). A further promising development is the presence of attempts within archaeology to develop bespoke solutions for addressing specific kinds of missing data problems. For example, several correction formulas have been proposed for the analyses of time-frequency data (Bluhm and Surovell, 2019; Surovell et al., 2009), while the location of missing sites have been imputed before pursuing more complex analyses based on locational properties of known sites (Bevan and Wilson, 2013) or the presence of infrastructure and road networks (Priβ et al., 2025). While still small in number, these endeavours showcase a fruitful area of development in statistical inference tailored to challenges specific to our field.

3.2. Measurement error

Measurement error represents another class of inferential challenge that shares some commonalities with the missing data problem discussed in the previous section. Despite its ubiquity across different domains of archaeological research, the use of formal statistical methods is even more limited in this case. The consequences of measurement error are, however, widely known in the statistical literature: 1) they introduce biases in the parameter estimates, 2) they reduce statistical power, hindering our ability to infer relationships between variables, and 3) they limit our ability to detect features in graphic analysis (Carroll et al., 2006). As noted above, part of the problem is that measurement error pertains to different stages of archaeological inference, and successful integration in the inferential process requires proper quantification and reporting in the first place.

Formal methods for quantifying measurement errors differ widely across the field. In strongly lab-based areas of applications such as ^{14}C dating (Scott et al., 2007), spectrometry (Drake et al., 2022), or stable isotope analyses (Coplen, 2011), established procedures are available so that errors can be formally quantified. However, the extent of appropriate reporting of measurement errors varies considerably, with systematic surveys in some research areas showing that available information is often not shared despite calls for better practices (Jardine and Cunjak, 2005; Johnson et al., 2024; Millard, 2014; Szpak et al., 2017; Vanderplicht and Hogg, 2006). Quantification of measurement error in research areas based on metric data typically consists of calculating intra- and inter-observer measurement error (Lyman and VanPool, 2009), although similar procedures have been implemented in field surveys (Hawkins et al., 2003) and explored in typological classification (Whittaker et al., 1998). Inter- and intra-observer errors are, however, not always reported in final studies as their primary objective is often perceived to be a means for reassuring that measurement error is small and negligible rather than offering grounds for subsequent integration of the calculated errors in statistical modelling.

The quantification of measurement errors associated with categorical and ordinal variables (i.e. misclassifications) is particularly challenging, as effectively, they require the assignment of probability values for each possible level of observation. In many practical applications, however, uncertainty is expressed by generating new categorical levels rather than assigning specific probability values. For example, biological sex estimates often include additional levels expressing different levels of uncertainty, such as ‘probably female’ or ‘indeterminate’(Buikstra and Ubelaker, 1994). Typo-chronological uncertainty can adopt a similar procedure, although in most cases in an unsystematic way (but see Nakoinz, 2012). Bevan and colleagues (Bevan et al., 2012) discuss potential approaches for expressing typo-chronological uncertainties, opting for a solution that requires experts to assign a ‘percentage of confidence’ to relative dating levels (e.g. “70 % Hellenistic, 0 % Early Roman, 0 % Middle Roman and 30 % Late Roman”). Notwithstanding that, such an approach would theoretically need to account for an additional layer of uncertainty (i.e. probabilistic estimates for each dating level can itself be associated with uncertainty, e.g. 60–80 % Hellenistic and 20–40 % Late Roman), as well as inter- and intra-observer errors, the benefits of such venture is not limited to the possibility of a more straightforward integration to subsequent analysis. As noted by Bevan and colleagues, examining *how* these subjectively assigned probability values are distributed can reveal key insights on diagnosticity and potential directional biases in misclassification. Yet these practices remain rare, perhaps due to the notion that these additional efforts do not bring much benefit in the end and require experts to put numbers into something that is fundamentally fuzzy.

In the statistical literature, measurement errors are often classified into two types. In the case of *classical measurement error*, the observed value $\mathrm{W_i}$ of sample *i* is defined as

$$\mathrm{W_i} = \mathrm{X_i} + \mathrm{U_i} \qquad [1]$$

where $\mathrm{X_i}$ is the unknown ‘true’ value of *i* plus a random error $\mathrm{U_i}$. Typically, $\mathrm{U_i}$ is described as a Gaussian with a mean of 0 (i.e. the error is unbiased) and a standard deviation expressing the potential magnitude of the measurement error. Classical measurement errors are thus suitable in describing limited instrumental precision and are assumed to be non-differential, with $\mathrm{U_i}$ independent from $\mathrm{X_i}$. Notably, the variability of the observed values $\mathrm{W_i}$ is larger than the variability of the true value $\mathrm{X_i}$. In some cases, however, it is more appropriate to express the relationship between these terms as follows:

$$\mathrm{X_i} = \mathrm{W_i} + \mathrm{U_i} \qquad [2]$$

Equation [2] describes what is commonly referred to as *Berkson measurement error* (Berkson, 1950). Here, the measurement error is independent of the observed values, and the variability of $\mathrm{W_i}$ is smaller than the variability of the actual value $\mathrm{X_i}$. An archaeological example of this type of measurement error might be the chronological assignment of samples to a particular time interval (e.g. 900-700 BC) based on their affiliation to a particular cultural period (e.g. ‘Geometric’). While the true date of objects affiliated with such cultural period would differ (i.e., each sample will have a different $\mathrm{X_i}$), all samples would be described identically (i.e., they would have the same $\mathrm{W_i}$). Both forms of uncertainty are widely present in archaeology.

As noted above, measurement errors are often not adequately accounted for in many archaeological applications, even when their quantification is provided. The most common example pertains to using archaeological dates as a variable in a statistical model. Regression analyses often employ mid-points of typo-chronologically defined time intervals or descriptive measures of central tendencies such as mean and median. Examples of this practice are relatively common and can be found in studies focused on inferring rates of dispersal (Pinhasi et al., 2005) or crop domestication (Fuller et al., 2012), but also in estimating the origination date of a particular technology (Bebber and Key, 2022) adapting methods developed in palaeontology. In many (but not all) cases, the temporal scope is sufficiently broad to justify the assumption that the effect of individual measurement error is negligible. Fig. 2 provides a counterexample and illustrates the potential danger of this practice when, given particular conditions (e.g. the presence of a calibration plateau), parameter estimates (e.g. the slope of a regression line) can be biased.

The issue of disregarding measurement error in chronological analyses has long been argued by archaeologists (Gkiasta et al., 2003) and has typically been approached by either 1) removing samples with more significant measurement errors, 2) employing a resampling-based approach, 3) directly integrating error terms in the statistical model. Removing selective samples with large measurement errors works under the premise that this would lead to $\mathrm{W_i} \simeq \mathrm{X_i}$ at the cost of reduced statistical power. However, if $\mathrm{U_i}$ is conditional to variables influencing $\mathrm{X_i}$, the artificial creation of missing data may introduce biases in estimated parameters. Resampling-based methods consist of replicating the statistical procedure (e.g. fitting a linear model) *n* times, using random sets of observations obtained by sampling from the probability distribution at each iteration describing the uncertainty of each data point. The approach has been used in the context of regression analyses involving radiocarbon dates (Fort, 2022; Gangal et al., 2014; Gkiasta et al., 2003; Riris and Silva, 2021), but also to carry out time-frequency analyses

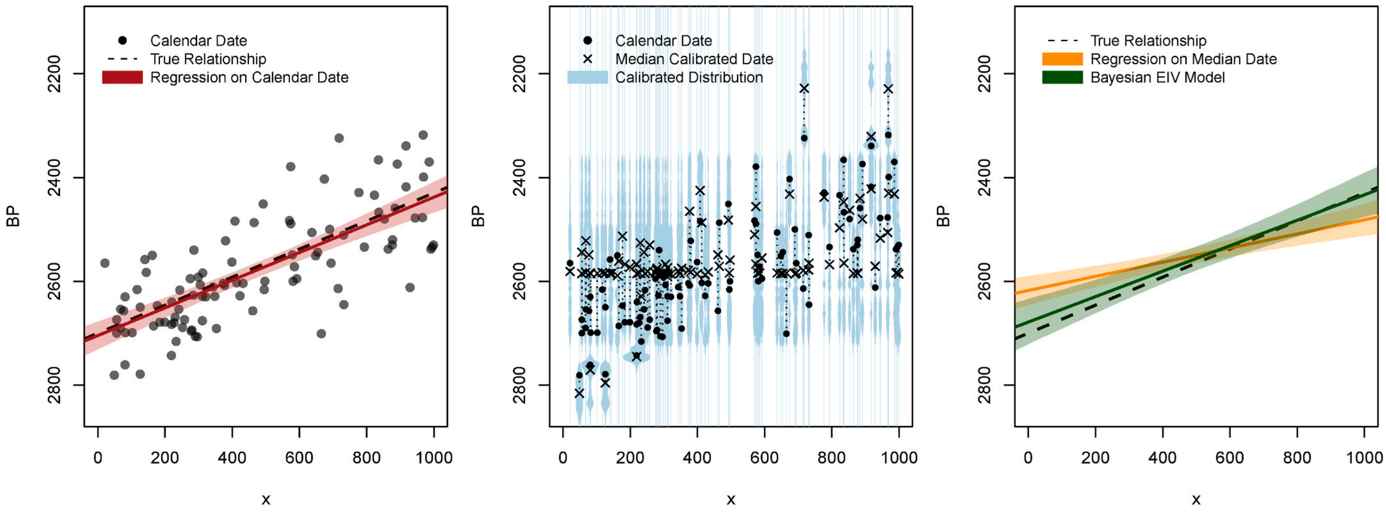

Fig. 2. A simulated dataset showing how ignoring measurement error can introduce biases in regression parameter estimates: (*left*) unbiased regression estimates of calendar dates against a predictor variable *x* show correct recovery of the 'true' parameter; (*middle*) comparison between calendar date, calibrated distribution, and median calibrated distribution on the same datasets showing the impact of calibration plateau; (*right*) regression estimates based on median calibrated dates fail to recover the 'true' relationship between *x* and the dates whilst a Bayesian EIV model successfully recovers the slope parameter.

based on archaeological periodisations (Baxter and Cool, 2016; Crema, 2012; Orton et al., 2017), and estimates of extinction dates in palaeontology (Herrando-Pérez and Saltré, 2024). Applications outside chronology are less common but do exist. For example, a recent paper by Lewis introduces a similar resampling approach to measurement error in digital elevation models in the context of least-cost path analysis (Lewis, 2021). An important but sometimes under-discussed step of these resampling methods is how the *n* results of the statistical analyses are aggregated into a single estimate with its error term. The exact procedure to achieve this varies between applications. In some areas of applications, such as time-frequency analysis, individual error terms are ignored, and as such, what is reported is a descriptive rather than an inferential statistic (i.e. sampling error is not accounted for; see Crema, 2024 for further discussion). In other applications, error terms of each of the *n* regression parameters are aggregated (Riris and Silva, 2021).

A more direct integration of the measurement error can be achieved using what is often referred to as Bayesian error-in-variable (EIV) models (sometimes also referred to as measurement error models), a particular form of multilevel model where each observation is described by a statistical distribution portraying its uncertainty. While this terminology is not used, the formal inclusion of measurement errors in Bayesian statistical models has a long history of application in the analysis of radiocarbon dates (Buck et al., 1996) and is implemented in widely used software packages such as *OxCal* (Bronk Ramsey, 1995). Bayesian EIV models provide both posterior estimates of higher-level parameters of interest (e.g. the start and end dates of an archaeological phase) as well as of individual observations, providing for the latter a reduction of the original measurement error of each sample informed by the model and the entire dataset. The highly specialised nature of most software applications dedicated to the Bayesian analysis of radiocarbon dates confines the application of these models to a limited set of research contexts (i.e. phase and age-depth models) despite its potential to employ these for a wider range of archaeological questions. For example, Crema (2024) model the diffusion of farming and fluctuations in the relative proportion of cremation and inhumation burial by examining binary (presence/absence) data points associated with radiocarbon dates and employ a Bayesian EIV model to account for the measurement errors associated with the latter. Some recent works on time-frequency analyses of radiocarbon dates similarly employ an EIV model under the hood, extending the range of applications beyond phase and age-depth models (Crema and Shoda, 2021; Heaton, 2022; Price et al., 2021). Applications of these models are certainly not limited to radiocarbon dates or continuous measurements. Groβ (Groβ, 2016; see also Rosenstock et al., 2019) used Bayesian EIV models to account for misclassification probabilities of biological sex (using Beta distribution to describe uncertain levels such as 'Female?' and 'Indeterminate') and chronological uncertainty associated with archaeological periodisation (using a uniform distribution bounded by the presumed start and end date of each period) when investigating long-term spatiotemporal variations in human stature. Despite the clear advantages offered by EIV models, their applications in archaeology are currently limited, owing in part to the lack of off-the-shelf software applications. Still, they offer the potential to become a standard toolkit for many statistical applications in archaeology where these form of measurements errors are common.

4. Simulations and generative inference

A core methodological foundation underpinning all the inferential approaches discussed so far is the reliance on the likelihood function, which requires the ability to mathematically define the probability of observing a specific outcome for a particular model parametrisation (e.g. what is the probability that 7 out of 12 burials had grave goods under a binomial distribution with $p = 0.5$ and $n = 12$). Both frequentist and Bayesian paradigms are centred on our ability to describe observed data using probability distributions, and likelihood functions to test particular hypotheses or estimate model parameters. While the level of abstraction offered by these models allows us to use the same inferential engine to investigate a wide range of empirical phenomena, they inevitably offer a relatively weak and indirect link between theory and data (Deffner et al., 2024).

We see similar separation between theory and data on other kinds of archaeological modelling as well, most notably in the applications of agent-based simulations (Lake, 2013; Romanowska and Wren, 2021). Here, nearly three decades of research have shown little analytical sophistication in comparing model expectations to empirical evidence. While the lack of effort is in part explained by the fact that many of these models were designed as theory-building tools (and hence never intended to be tested directly against evidence and designed more to explore the cumulative consequences of different assumptions), it is safe to state that even empirically and contextually grounded simulations are often compared to archaeological evidence in loose qualitative terms. Thus, we have, on the one hand, a series of robust inferential tools that can be used to describe empirical evidence in terms of abstract and highly generalisable conditional probability distributions and, on the other hand, mechanistic models that can formally represent assumed behavioural and causal generative processes of observed patterns, but can only indirectly be referenced during data analysis.

Approximate Bayesian Computation (ABC), and more broadly, *generative inference* (Kandler and Powell, 2018), provides an intuitive bridge that connects more directly formal models to empirical data. In its simplest form, the inferential machinery of ABC consists of a simple idea (see Fig. 3).

1. Define a data-generating simulation model $\mathscr{M}$, tuned with a set of *k* parameters Θ_1, Θ_2,. Θ_k
2. Associate each of the *k* parameters with a probability distribution defining their *prior.*
3. Sample a set of possible parameter values from the prior distributions of Θ_1, Θ_2,. Θ_k.
4. Feed the parameters obtained in step 3 to $\mathscr{M}$, and generate an output that can be compared to the observed data.
5. Repeat steps 3–4 *n* times to obtain a distribution of outputs.

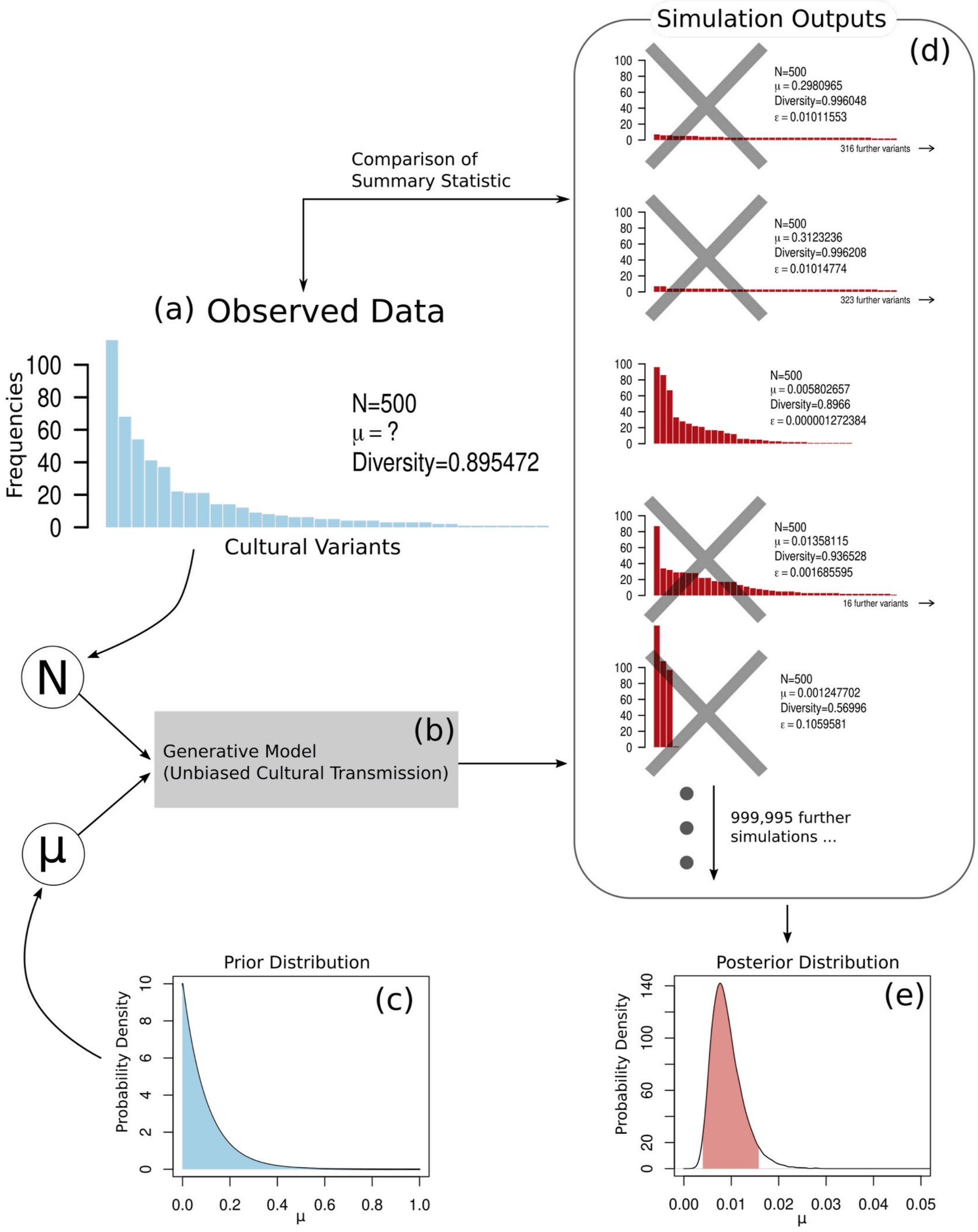

Fig. 3. An illustration of the basic workflow of Approximate Bayesian Computation using the rejection algorithm. An observed frequency data of cultural variants (a) is described by a summary statistic (diversity index) and assumed to result from an unbiassed transmission process with an unknown innovation rate μ (Kandler and Crema, 2019). A generative model (b) is used to generate artificial data with parameters either derived from the data (population size *N*) or a prior distribution (c). The summary statistics of the resulting simulation outputs (d) are compared against the observed values to obtain a distance value (ε); simulations with ε exceeding the tolerance threshold (here set to 0.00001) are rejected (d) and the distribution of μ values that were used to generate the remaining simulations becomes the posterior of μ (e).

6. Measure the distance ε between the simulation output and the observed data.
7. Discard parameter values where ε is larger than a pre-defined tolerance threshold τ; the distributions of the remaining sampled values of Θ_1, Θ_2,. Θ_k are the approximate *posterior* of the model parameters.

Core ideas of ABC date back to the 1980s (Rubin, 1984), although its modern reincarnation stems from applications in population genetics in the early 2000s (Beaumont et al., 2002), and since then, it has been a successful inferential tool used in a wide range of disciplines (see Sisson et al., 2018 for a review). Two decades of applications have also led to considerable methodological improvement with improved precision and reduced computational costs compared to the original *rejection algorithm* described above.

The appeal of an inferential tool that bypasses the need for a likelihood function and benefits from the flexibility of computational models in defining processes and relationships between variables, has not gone unnoticed in archaeology. Early applications include Tsutaya and Yoneda's (2013) study on estimating weaning age from Nitrogen Isotope Ratios of bone collagen, Porčić and Nikolić's (2016) demographic reconstruction of Lepenski Vir, and several studies focusing on cultural transmission models of material culture (Crema et al., 2014; Kandler and Shennan, 2015; Kovacevic et al., 2015). The availability of a rich body of formal theoretical models developed in cultural evolutionary science over four decades (Boyd and Richerson, 1985; Cavalli-Sforza and Feldman, 1981) have led to particularly prolific areas of applications in archaeology (Carrignon et al., 2020, 2023; Crema et al., 2014, 2016; Kandler and Shennan, 2015) and adjacent fields (Carrignon et al., 2019; Pagel et al., 2019; Youngblood, 2019; Youngblood et al., 2023; Youngblood and Lahti, 2022). Other applications of ABC in archaeology include the identification of points of origin of diffusion of farming (Cortell-Nicolau et al., 2021) and demographic trajectories as inferred from time-frequency of radiocarbon dates (Cortell-Nicolau et al., 2025; DiNapoli et al., 2021).

Typical applications of ABC require substantial effort on the researchers' part, who can rarely benefit from complete off-the-shelf solutions and often have to develop, on top of the core simulation model, other aspects of the computational pipeline (e.g. the specific algorithm to obtain good posterior samples with reduced computational cost). Despite such challenges, the moderate success of ABC in archaeology and anthropology testifies to the potential of this approach. In some research areas, such as cultural evolutionary studies, the approach is increasingly regarded as a key inferential tool for observational studies (Deffner et al., 2024; Kandler and Powell, 2018). Yet, ten years of applications of ABC in archaeology and anthropology have also identified several key issues that effectively limit its use. Firstly, ABC typically requires an extremely large number of simulation runs (typically in the order of 10^6) to achieve satisfactory precision and accuracy in parameter estimates. The resulting computational costs are onerous, requiring each simulation run to be completed in no more than a few seconds, a limit that is easily surpassed in most agent based models. While the development of efficient alternatives to the rejection algorithm has reduced the number of required simulation runs, as it stands, ABC is simply unfeasible when the objective is to fit complex models with more extensive runtimes. If likelihood-based alternatives are available, researchers should not use ABC as it offers reduced precision with substantial computational costs (see Crema, 2022 for an archaeological example and comparison).

Secondly, ABC hinges on the ability of simulation models to generate the observed data. In the majority of applications, however, the best we can achieve is to produce outputs that are close *enough* to the observed data. Determining a suitable tolerance threshold (see point 7 above) is, however, not trivial, and in many cases, researchers have opted for an approach where the posteriors are obtained by identifying the simulation runs with the closest fit to the data. Furthermore, in many circumstances, simulation outputs and observations are compared using (often insufficient) summary statistics. While such an approach offers more flexibility, it inevitably entails a substantial loss of information, which can lead to wider posterior intervals.

A third practical limitation of simulation-based generative inference stems from the fact that, in some cases, the initial conditions of the system of interest must be fed into the model. While certain processes entailing information loss, such as taphonomy and time-averaging, can be modelled in the simulation itself, allowing comparison to observed data similarly affected by these processes, the uncertainties associated with the initial conditions are much harder to account for (Crema et al., 2016). These challenges inevitably narrow the scope of application of generative inference compared to more conventional likelihood-based statistical models discussed above.

5. Discussion and conclusion

Archaeological data offers some unique challenges for quantitative analysis, to the extent that some may perceive any endeavour in applying statistical analysis as futile. After all, the observations we have in hand are indirect proxies of past behaviour in small samples, biased by the idiosyncrasies of highly diverse retrieval and sampling practices, and often limited by the uncertainty associated with large measurement errors and missing data. This perceived disciplinary exceptionalism, however, ignores advances made in parallel fields of studies where similar problems have been tackled over the last few decades. While the borrowing of statistical techniques from other disciplines is sometimes viewed with suspicion and does raise concerns about potential misuse and inappropriate adaptation of underlying theoretical tenets, the approaches presented here offer general principles that are flexible enough to incorporate diverse bodies of theories. Naturally, as with any other statistical techniques, the methods described here are not immune to such abuses and misinterpretations, but they offer a significant advance to alternatives and standard employed currently in the field. Indeed, one could argue that choosing a standard linear regression ignoring sample interdependency or measurement errors is *de facto* a form of statistical abuse that just happens to be so common to be ignored. There are better ways to do quantitative archaeology.

It is difficult to determine to what extent the conclusions of past archaeological investigations were biased due to the mishandling of factors such as sample interdependence, missing data, or measurement error. The 'moralising god' case discussed in section 3.1 is sadly still a rare case of a success story where the open science practice embraced by the original authors is what allowed other scholars to explore the consequences of these biasing factors. The relevance of computational reproducibility in archaeology is hence pivotal (Marwick et al., 2017), not just to promote the application of increasingly complex techniques and allow for transparency in communicating key aspects of the research pipeline but also for the field to cumulatively progress from previous work, addressing issues and challenges as we move forward. Many (although sadly not all) of the archaeological applications of the three techniques discussed here were published following the principles of open science — readers interested in understanding and exploring these techniques will have the opportunity to gain information that just a decade ago would have been inaccessible.

Among the techniques reviewed here, multilevel models will likely become part of the standard tool in quantitative archaeology. Sister disciplines such as biological anthropology and ecology routinely employ these models, and a large number of software applications offer off-the-shelves solutions (Fernée and Trimmis, 2021; McCoach et al., 2018). Given the advantages they offer, multilevel models ought to become a standard inferential tool, particularly in regional and cross-regional studies that draw conclusions from multiple sites. The shift towards multilevel models does not only address potential challenges such as sampling imbalance but also problematic assumptions embedded in more conventional methods. For example, by modelling the variability in the relationship between a predictor and a response at the group level, multilevel models provide a solution to the long-standing problem of environmental determinism, effectively addressing an ‘impoverished representation of reality’ where a ‘single general relationship across time and space’ is assumed to take place (Jones, 1991, p. 8). However, the most significant potential offered by multilevel models, particularly within a Bayesian framework, is the opportunity to construct complex models tailored to specific needs. Some examples of archaeological applications are moving towards this direction, exploring relationships between compositional variances and covariates rather than modelling variation in modes or means (Vieri et al., 2025), or combining mixing models to explore intersite variations in faunal sex composition using morphometric data (Wolfhagen, 2024). Developing these custom models requires both mathematical and computational skills, but in some research areas, such as mixing models for biotracers (Stock et al., 2018) and chronological modelling (Lanos and Philippe, 2018), dedicated software applications have been published, enabling the opportunities to widen the application of multilevel models further.

Still, some areas of applications remain relatively underexplored. Bayesian EIV models are effectively a type of multilevel model that can overcome the challenges imposed by measurement errors but have been almost exclusively used to analyse radiocarbon dates. The investment in these applications was promoted by the necessity to address chronological uncertainty but also by the opportunities offered by a type of data where formal measurements of error are comparatively straightforward to measure and obtain. Notwithstanding the challenges imposed by formally measuring time from culturally diagnostic features (but see e.g. Carleton et al., 2023), the integration of other forms of chronologies that are archaeologically far more common is a necessary step for future work of synthesis of legacy datasets.

But the methods reviewed in this paper are not straightforward to implement. They require solid understanding of statistical theory and the ability to adapt research design and models to specific question and datasets. There is no easy way out, and the range of case studies where basic text-book statistical tools are appropriate is perhaps much smaller than we wish. Applyng these techniques is a challenging and time-consuming task, but a concerted effort can provide benefits to the entire discpline, offering examples and alternatives rather than narrowly prescribed simple instructions that can lead to biased inference.

A fundamental tool that can help researchers develop a suitable research design and statistical model is to employ simulations to generate artificial dataset that can then be used to evaluate the robustness of a particular approach. Buck and Meson (2015) use the term ‘What if experiments’ to refer to the practice of using simulated data to evaluate sampling strategies or consider the implications of particular assumptions in the context of Bayesian chronological modelling. Such an approach can, and should be, part of a standard toolkit in quantitative archaeology. It can provide insights on how or model can perform, but also elucidate the relationship between key concepts such as statistical significance, precision of the parameter estimate, effect size, and sample size. Much of the pit-falls in the misinterpretation of concepts like p-values could be avoided with a better grasp on how these factors are interlinked.

However, these advanced statistical techniques will just offer a glorified detection of correlative patterns in the archaeological record if we do not associate them with a proper causal inference framework. Archaeologists need to be simultaneously more conscious about the pitfalls of correlative models but also move beyond the ’*correlation doesn’t equal causation*’ adage. As a field that attempts to explain past human behaviour, we rarely have the opportunity to design controlled experiments and, in the great majority of cases, rely on observational studies. Yet, most regression-based analyses either opt for assessing sequentially single predictor variables, fit a model including every possible covariate and interpreting individual coefficients in causal terms, or employ some variable selection algorithms such as stepwise AIC (Akaike Information Criteria) and *then* interpret the resulting coefficients. All these approaches have implications in terms of causal inference. Examining individual covariates sequentially ignores the potential effect of confounder variables while including all terms will suffer from overfitting and potential biases introduced by so-called “colliders”(Cole et al., 2010). Both confounders and colliders can lead to biased estimates in regression models and can even produce parameter estimates suggesting reversed relationships. The widely adopted practice of fitting all potential causal variables of interest in a single statistical model and interpreting the resulting coefficients (explicitly or implicity) in causal terms will result in what some refer to as the ‘Table 2 fallacy’, and should be avoided (Westreich and Greenland, 2013). Invoking ’*correlation doesn’t equal causation*’ after committing this fallacy and inviting caution in interpreting model outputs in causal terms is a confusing, misleading, and lazy malpractice at best. The use of AIC and related methods is becoming common in archaeology, and while they do help avoid over-fitting, they are designed with out-of-sample predictive accuracy in mind and hence are not able to suggest what the appropriate set of covariates to be included in a statistical model where causal interpretations are sought. The problem of appropriate variable selection and its relation to causal inference is virtually undiscussed in archaeology, but problems and solutions do exist and are being increasingly employed in fields such as epidemiology, ecology, and evolutionary anthropology (Deffner et al., 2024; Rothman and Greenland, 2005; Shipley, 2016).

While an extensive review of the causal inference literature is beyond the scope of this review (interested readers should consult McElreath, 2020), it is worth highlighting a key element required to pursue such an endeavour. Currently, the most common approach for determining an appropriate set of control variables to be included in a statistical model when determining the causal effect of a particular exposure variable is to construct a direct acyclic graph (DAG). A DAG is a graphical representation of causal assumptions (Bulbulia, 2024; Rohrer, 2018) that helps determine which variables to control and which ones not to control, but also to determine whether a causal inference is even possible with the data available, potentially informing research design. Importantly, there is no reason to assume that for a given phenomenon of interest, there is only one possible DAG. Different theories and assumptions will lead to different presumed causal relationships, which in turn may or may not require different research designs and statistical models.

Both DAGs and the generative inference framework discussed in section 4 thus require us to be explicit about processes behind the observed patterns in the archaeological record *before* carrying out statistical analyses and interpreting its outputs. This requirement will help establish more direct link between theory, hypothesis, and quantitative analysis, and can serve as valuable heuristic tools even when they are not supported by empirical evidence. Importantly, they also help us avoid the temptation of building *post-hoc accommodative arguments* (Binford, 1981) or pursue *HARKing* (“Hypothesising after the results are known”; Kerr, 1998). In a recent critical reflection piece, Michael Smith (2023) distinguishes between “internal” and “external” arguments, with the former being the formal model itself and the latter being its validation through a comparison to external data. As noted by Smith, there is a long tradition of archaeological work on the former, most recently with the increasing use of agent-based simulations (Lake, 2013; Romanowska and Wren, 2021), but the focus on the latter has been lagging behind. Without external validation, a formal model is just an opinion, an internally and logically consistent one, but nonetheless just an untested claim about our past. Advances in statistical methods are now providing the necessary tools to test explicitly our theories and move from correlation to causation whilst accounting for the complexity and limitations of our data. Notwithstanding the new challenges we will need to address in pursuing this endeavour, there are no excuses for not stepping up.

Data availability statement

All R scripts required to generate Figs. 1–3 are available on the following GitHub repository: https://github.com/ercrema/statistical_modelling_review and archived on: https://doi.org/10.5281/zenodo.14800851.

Reproducible results

The Associate Editor for Reproducibility downloaded all materials and could reproduce the results presented by the authors.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

I am grateful to Emma Pomeroy and Tamsin O'Connell for providing some comments and feedbacks on specific aspects of the manuscript, and to the three anonymous reviewers for providing key suggestions that helped improved the manuscript.

References

Aldenderfer, M., 1998. Quantitative methods in archaeology: a review of recent trends and developments. J. Archaeol. Res. 6, 91–120.

Allison, P.D., 2002. Missing Data, Quantitative Applications in the Social Sciences. SAGE Publications, Thousand Oaks, CA.

Alonso-Llamazares, C., Lopez, B., Pardiñas, A., 2022. Sex differences in the distribution of entheseal changes: meta-analysis of published evidence and its use in Bayesian paleopathological modeling. Am. J. Biol. Anthropol. 177, 249–265.

Altschul, J.H., Kintigh, K.W., Klein, T.H., Doelle, W.H., Hays-Gilpin, K.A., Herr, S.A., Kohler, T.A., Mills, B.J., Montgomery, L.M., Nelson, M.C., Ortman, S.G., Parker, J.N., Peeples, M.A., Sabloff, J.A., 2018. Fostering collaborative synthetic research in archaeology. Adv. Archaeol. Pract. 6, 19–29.

Ammerman, A., 1992. Taking stock of quantitative archaeology. Annu. Rev. Anthropol. 21, 231–249.

Banks, W.E., Bertran, P., Ducasse, S., Klaric, L., Lanos, P., Renard, C., Mesa, M., 2019. An application of hierarchical Bayesian modeling to better constrain the chronologies of Upper Paleolithic archaeological cultures in France between ca. 32,000–21,000 calibrated years before present. Quat. Sci. Rev. 220, 188–214.

Banning, E.B., 2021. Sampled to death? The rise and fall of probability sampling in archaeology. Am. Antiq. 86, 43–60.

Barcelo, J.A., Bogdanovic, I. (Eds.), 2015. Mathematics and Archaeology. CRC Press, Bosa Roca.

Baxter, M., 2003. Statistics in Archaeology, Arnold Applications of Statistics. Hodder Arnold, London, England.

Baxter, M.J., Cool, H.E.M., 2016. Reinventing the wheel? Modelling temporal uncertainty with applications to brooch distributions in Roman Britain. J. Archaeol. Sci. 66, 120–127.

Beaumont, M.A., Zhang, W., Balding, D.J., 2002. Approximate bayesian computation in population genetics. Genetics 162, 2025–2035.

Bebber, M.R., Key, A.J.M., 2022. Optimal linear estimation (OLE) modeling supports early Holocene (9000–8000 RCYBP) copper tool production in North America. Am. Antiq. 87, 1–17.

Beheim, B., Atkinson, Q.D., Bulbulia, J., Gervais, W., Gray, R.D., Henrich, J., Lang, M., Monroe, M.W., Muthukrishna, M., Norenzayan, A., Purzycki, B.G., Shariff, A., Slingerland, E., Spicer, R., Willard, A.K., 2021. Treatment of missing data determined conclusions regarding moralizing gods. Nature 595, E29–E34.

Berkson, J., 1950. Are there two regressions? J. Am. Stat. Assoc. 45, 164–180.

Bevan, A., 2015. The data deluge. Antiquity 89, 1473–1484.

Bevan, A., Conolly, J., Hennig, C., Johnston, A., Quercia, A., Spencer, L., Vroom, J., 2012. Measuring chronological uncertainty in intensive survey finds. Archaeometry 55, 318–328.

Bevan, A., Cutler, B., Hennig, C., Yermeche, O., 2024. Cereal silo-pits, Agro-pastoral practices and social organisation in 19th century Algeria. Hum. Ecol. Interdiscip. J. 52, 497–513.

Bevan, A., Wilson, A., 2013. Models of settlement hierarchy based on partial evidence. J. Archaeol. Sci. 40, 2415–2427.

Binford, L., 1981. Bones: Ancient Men and Modern Myths. Academic Press, New York.

Bluhm, L.E., Surovell, T.A., 2019. Validation of a global model of taphonomic bias using geologic radiocarbon ages. Quat. Res. 91, 325–328.

Bordes, F., 1953. Essai de classification des industries moustérienne. Bull. Soc. Prehist. Fr. 50, 457–466.

Boyd, R., Richerson, P.J., 1985. Culture and the Evolutionary Process. University of Chicago Press, Chicago.

Brainerd, G.W., 1951. The place of chronological ordering in archaeological analysis. Am. Antiq. 16, 301–313.

Bronk Ramsey, C., 1995. Radiocarbon calibration and analysis of stratigraphy: the OxCal program. Radiocarbon 37, 425–430.

Buck, C.E., Cavanagh, W.G., Litton, C.D., 1996. Bayesian Approach to Interpreting Archaeological Data. Wiley, Chirchester.

Buck, C.E., Meson, B., 2015. On being a good bayesian. World Archaeol. 47, 567–584.

Buikstra, J.E., Ubelaker, D.H., 1994. In: Standards for Data Collection from Human Skeletal Remains, 44. Archaeological Survey Research, Fayetteville, Arkansas.

Bulbulia, J.A., 2024. Methods in causal inference. Part 1: causal diagrams and confounding. Evol. Hum. Sci. 6, e40.

Carleton, W.C., Klassen, S., Niles-Weed, J., Evans, D., Roberts, P., Groucutt, H.S., 2023. Bayesian regression versus machine learning for rapid age estimation of archaeological features identified with lidar at Angkor. Sci. Rep. 13, 17913.

Carlson, D.L., 2017. Cambridge Manuals in Archaeology: Quantitative Methods in Archaeology Using R. Cambridge University Press, Cambridge, England.

Carrignon, S., Alexander Bentley, R., O'Brien, M.J., 2023. Estimating two key dimensions of cultural transmission from archaeological data. J. Anthropol. Archaeol. 72, 101545.

Carrignon, S., Bentley, R.A., Ruck, D., 2019. Modelling rapid online cultural transmission: evaluating neutral models on Twitter data with approximate Bayesian computation. Palgrave Commun 5, 1–9.

Carrignon, S., Brughmans, T., Romanowska, I., 2020. Tableware trade in the Roman East: exploring cultural and economic transmission with agent-based modelling and approximate Bayesian computation. PLoS One 15, e0240414.

Carroll, R.J., Ruppert, D., Stefanski, L.A., Crainiceanu, C.M., 2006. Measurement Error in Nonlinear Models. CRC Press, London, England.

Cavalli-Sforza, L.L., Feldman, M.W., 1981. Cultural Transmission and Evolution: a Quantitative Approach. Princeton University Press.

Clarke, D., 1973. Archaeology: the loss of innocence. Antiquity 47, 6–18.

Cole, S.R., Platt, R.W., Schisterman, E.F., Chu, H., Westreich, D., Richardson, D., Poole, C., 2010. Illustrating bias due to conditioning on a collider. Int. J. Epidemiol. 39, 417–420.

Comer, J.A., Comer, D.C., Dumitru, I.A., Priebe, C.E., Patsolic, J.L., 2023. Sampling methods for archaeological predictive modeling: spatial autocorrelation and model performance. J. Archaeol. Sci. Rep 48, 103824.

Coplen, T.B., 2011. Guidelines and recommended terms for expression of stable-isotope-ratio and gas-ratio measurement results. Rapid Commun. Mass Spectrom. 25, 2538–2560.

Cortell-Nicolau, A., García-Puchol, O., Barrera-Cruz, M., García-Rivero, D., 2021. The spread of agriculture in Iberia through approximate Bayesian computation and Neolithic projectile tools. PLoS One 16, e0261813.

Cortell-Nicolau, A., Rivas, J., Crema, E.R., Shennan, S., García-Puchol, O., Kolář, J., Staniuk, R., Timpson, A., 2025. Demographic interactions between the last hunter-gatherers and the first farmers. Proc. Natl. Acad. Sci. USA 122, e2416221122.

Cowgill, G.L., 2015. Some things I hope you will find useful Even if statistics isn't your thing. Annu. Rev. Anthropol. 44, 1–14.

Crema, E.R., 2024. A Bayesian alternative for aoristic analyses in archaeology. Archaeometry. https://doi.org/10.1111/arcm.12984.

Crema, E.R., 2022. Statistical inference of prehistoric demography from frequency distributions of radiocarbon dates: a review and a guide for the perplexed. J. Archaeol. Method Theor 29, 1387–1418.

Crema, E.R., 2012. Modelling temporal uncertainty in archaeological analysis. J. Archaeol. Method Theor 19, 440–461.

Crema, E.R., Bloxam, A., Stevens, C.J., Vander Linden, M., 2024. Modelling diffusion of innovation curves using radiocarbon data. J. Archaeol. Sci. 165, 105962.

Crema, E.R., Edinborough, K., Kerig, T., Shennan, S.J., 2014. An approximate Bayesian computation approach for inferring patterns of cultural evolutionary change. J. Archaeol. Sci. 50, 160–170.

Crema, E.R., Kandler, A., Shennan, S., 2016. Revealing patterns of cultural transmission from frequency data: equilibrium and non-equilibrium assumptions. Sci. Rep. 6, 39122.

Crema, E.R., Shoda, S., 2021. A Bayesian approach for fitting and comparing demographic growth models of radiocarbon dates: a case study on the Jomon-Yayoi transition in Kyushu (Japan). PLoS One 16, e0251695.

Crema, E.R., Stevens, C.J., Shoda, S., 2022. Bayesian analyses of direct radiocarbon dates reveal geographic variations in the rate of rice farming dispersal in prehistoric Japan. Sci. Adv. 8, eadc9171.

Deffner, D., Fedorova, N., Andrews, J., McElreath, R., 2024. Bridging theory and data: a computational workflow for cultural evolution. Proc. Natl. Acad. Sci. USA 121, e2322887121.

DiNapoli, R.J., Crema, E.R., Lipo, C.P., Rieth, T.M., Hunt, T.L., 2021. Approximate Bayesian computation of radiocarbon and paleoenvironmental record shows population resilience on Rapa Nui (Easter Island). Nat. Commun. 12, 3939.

Drake, B.L., Mayer, E.G., Shugar, A., 2022. Uncertainty and pXRF measurements. In: Advances in Portable X-Ray Fluorescence Spectrometry. The Royal Society of Chemistry, Cambridge, pp. 447–489.

Drennan, R.D., 2009. Statistics for Archaeologists: a Common Sense Approach. Springer Science & Business Media.

Drennan, R.D., Earle, T., Feinman, G.M., Fletcher, R., Kolb, M.J., Peregrine, P., Peterson, C.E., Sinopoli, C., Smith, M.E., Smith, M.L., Stark, B.L., Stark, M.T., 2011. Comparative archaeology: a commitment to understanding variation. In: Smith, M.E. (Ed.), The Comparative Archaeology of Complex Societies. Cambridge University Press, Cambridge, pp. 1–3.

Fanta, V., Zouhar, J., Beneš, J., Bumerl, J., Sklenicka, P., 2020. How old are the towns and villages in Central Europe? Archaeological data reveal the size of bias in dating obtained from traditional historical sources. J. Archaeol. Sci. 113, 105044.

Fernée, C.L., Trimmis, K.P., 2021. Detecting variability: a study on the application of bayesian multilevel modelling to archaeological data. Evidence from the Neolithic Adriatic and the Bronze Age Aegean. J. Archaeol. Sci. 128, 105346.

Fletcher, M., Lock, G., 2005. Digging Numbers, second ed. Oxford University School of Archaeology Monograph. Oxford University School of Archaeology, Oxford, England.

Fort, J., 2022. Dispersal distances and cultural effects in the spread of the Neolithic along the northern Mediterranean coast. Archaeol. Anthropol. Sci. 14. https://doi.org/10.1007/s12520-022-01619-x.

Fuller, D.Q., Asouti, E., Purugganan, M.D., 2012. Cultivation as slow evolutionary entanglement: comparative data on rate and sequence of domestication. Veg. Hist. Archaeobotany 21, 131–145.

Gangal, K., Sarson, G.R., Shukurov, A., 2014. The near-eastern roots of the Neolithic in South Asia. PLoS One 9, e95714.

Gero, J.M., 2007. Honoring ambiguity/problematizing certitude. J. Archaeol. Method Theor 14, 311–327.

Gkiasta, M., Russell, T., Shennan, S., Steele, J., 2003. Neolithic transition in Europe: the radiocarbon record revisited. Antiquity 77, 45–62.

Gonzalez-Perez, C., Pereira-Fariña, M., Martín-Rodilla, P., Tobalina-Pulido, L., 2023. Dealing with vagueness in archaeological discourses. In: Discourse and Argumentation in Archaeology: Conceptual and Computational Approaches. Springer International Publishing, Cham, pp. 137–157.

Groß, M., 2016. Modeling body height in prehistory using a spatio-temporal Bayesian errors-in variables model. Adv. Stat. Anal. 100, 289–311.

Hawkins, A.L., Stewart, S.T., Banning, E.B., 2003. Interobserver bias in enumerated data from archaeological survey. J. Archaeol. Sci. 30, 1503–1512.

Heaton, T.J., 2022. Non-parametric calibration of multiple related radiocarbon determinations and their calendar age summarisation. J. R. Stat. Soc. Ser. C Appl. Stat. 71, 1918–1956.

Heegaard, E., Birks, H.J.B., Telford, R.J., 2005. Relationships between calibrated ages and depth in stratigraphical sequences: an estimation procedure by mixed-effect regression. Holocene 15, 612–618.

Heizer, R.F., Cooke, S.F. (Eds.), 1960. The Application of Quantiative Methods in Archaeology. Quadrangle Books, Chicago.

Herrando-Pérez, S., Saltré, F., 2024. Estimating extinction time using radiocarbon dates. Quat. Geochronol. 79, 101489.

Jardine, T.D., Cunjak, R.A., 2005. Analytical error in stable isotope ecology. Oecologia 144, 528–533.

Johnson, K., Quinn, C.P., Goodale, N., Conrey, R., 2024. Best practices for publishing pXRF analyses. Adv. Archaeol. Pr. 12, 156–162.

Jones, K., 1991. Multi-level models for geographical research. Concepts and Techniques in Modern Geography, 54. Environmental Publications, Norwich, England.

Kandler, A., Crema, E.R., 2019. Analysing cultural frequency data: neutral theory and beyond. In: Prentiss, A.M. (Ed.), Handbook of Evolutionary Research in Archaeology. Springer International Publishing, Cham, pp. 83–108.

Kandler, A., Powell, A., 2018. Generative inference for cultural evolution. Phil. Trans. Biol. Sci. 373, 20170056.

Kandler, A., Shennan, S., 2015. A generative inference framework for analysing patterns of cultural change in sparse population data with evidence for fashion trends in LBK culture. J. R. Soc. Interface 12, 20150905.

Kerr, N.L., 1998. HARKing: hypothesizing after the results are known. Pers. Soc. Psychol. Rev. 2, 196–217.

Key, A., Roberts, D., Jarić, I., 2021. Reconstructing the full temporal range of archaeological phenomena from sparse data. J. Archaeol. Sci. 135, 105479.

Kintigh, K.W., Altschul, J.H., Beaudry, M.C., Drennan, R.D., Kinzig, A.P., Kohler, T.A., Limp, W.F., Maschner, H.D.G., Michener, W.K., Pauketat, T.R., Peregrine, P., Sabloff, J.A., Wilkinson, T.J., Wright, H.T., Zeder, M.A., 2014. Grand challenges for archaeology. Am. Antiq. 79, 5–24.

Kohler, T.A., Bogaard, A., Ortman, S.G., Crema, E.R., Chirikure, S., Cruz, P., Green, A., Kerig, T., McCoy, M.D., Munson, J., Petrie, C., Thompson, A.E., Birch, J., Cervantes Quequezana, G., Feinman, G.M., Fochesato, M., Gronenborn, D., Hamerow, H., Jin, G., Lawrence, D., Roscoe, P.B., Rosenstock, E., Erny, G.K., Kim, H., Ohlrau, R., Hanson, J.W., Fargher Navarro, L., Pailes, M., 2025. Economic inequality is fueled by population scale, land-limited production, and settlement hierarchies across the archaeological record. Proc. Natl. Acad. Sci. USA 122, e2400691122.

Kovacevic, M., Shennan, S., Vanhaeren, M., d'Errico, F., Thomas, M.G., 2015. Simulating geographical variation in material culture: were early modern humans in Europe ethnically structured? In: Mesoudi, A., Aoki, K. (Eds.), Learning Strategies and Cultural Evolution During the Palaeolithic, Replacement of Neanderthals by Modern Humans Series. Springer, Japan, pp. 103–120.

Kristiansen, K., 2021. Towards a new paradigm? The third science revolution and its possible consequences in archaeology. Curr. Swed. Archaeol. 22, 11–34.

Lake, M.W., 2013. Trends in archaeological simulation. J. Archaeol. Method Theor 21, 258–287.

Lanos, P., Philippe, A., 2018. Event date model: a robust Bayesian tool for chronology building. Commun. Stat. Appl. Methods 25, 131–157.

Leipe, C., Long, T., Sergusheva, E.A., Wagner, M., Tarasov, P.E., 2019. Discontinuous spread of millet agriculture in eastern Asia and prehistoric population dynamics. Sci. Adv. 5, eaax6225.

Leonard, R.D., Jones, G.T. (Eds.), 1989. Quantifying Diversity in Archaeology, New Directions in Archaeology. Cambridge University Press, Cambridge, England.

Lewis, J., 2021. Probabilistic modelling for incorporating uncertainty in least cost path results: a postdictive Roman road case study. J. Archaeol. Method Theor 28, 911–924.

Lyman, R.L., VanPool, T.L., 2009. Metric data in archaeology: a study of intra-analyst and inter-analyst variation. Am. Antiq. 74, 485–504.

Marwick, B., d'Alpoim Guedes, J., Barton, C.M., Bates, L.A., Baxter, M., Bevan, A., Bollwerk, E.A., Bocinsky, R.K., Brughmans, T., Carter, A.K., Conrad, C., Contreras, D. A., Costa, S., Crema, E.R., Daggett, A., Davies, B., Drake, B.L., Dye, T.S., France, P., Fullagar, R., Giusti, D., Graham, S., Harris, M.D., Hawks, J., Health, S., Huffer, D., Kansa, E.C., Kansa, S.W., Madsen, M.E., Melcher, J., Negre, J., Neiman, F.D., Opitz, R., Orton, D.C., Przstupa, P., Raviele, M., Riel-Savatore, J., Riris, P., Romanowska, I., Smith, J., Strupler, N., Ullah, I.I., Van Vlack, H.G., VanValkenburgh, N., Watrall, E.C., Webster, C., Wells, J., Winters, J., Wren, C.D., 2017. Open science in archaeology. SAA Archaeol. Rec. 17, 8–14.

McCoach, D.B., Rifenbark, G.G., Newton, S.D., Li, X., Kooken, J., Yomtov, D., Gambino, A.J., Bellara, A., 2018. Does the package matter? A comparison of five common multilevel modeling software packages. J. Educ. Behav. Stat. 43, 594–627.

McCullagh, P., 2002. What is a statistical model? Ann. Stat. 30, 1225–1310.

McElreath, R., 2020. Statistical Rethinking: a Bayesian Course with Examples in R and Stan, second ed. CRC Press.

Millard, A.R., 2014. Conventions for reporting radiocarbon determinations. Radiocarbon 56, 555–559.

Nakoinz, O., 2012. Datierungskodierung und chronologische Inferenz – techniken zum Umgang mit unscharfen chronologischen Informationen. Praehistorische Z. 87. https://doi.org/10.1515/pz-2012-0010.

Orton, C., 2012. Cambridge Manuals in Archaeology: Sampling in Archaeology. Cambridge University Press, Cambridge, England.

Orton, D., Morris, J., Pipe, A., 2017. Catch per unit research effort: sampling intensity, chronological uncertainty, and the onset of marine fish consumption in historic London. Open Quat. 3. https://doi.org/10.5334/oq.29.

Otárola-Castillo, E.R., Torquato, M.G., 2018. Bayesian statistics in archaeology. Annu. Rev. Anthropol. 47, 435–453.

Pagel, M., Beaumont, M., Meade, A., Verkerk, A., Calude, A., 2019. Dominant words rise to the top by positive frequency-dependent selection. Proc. Natl. Acad. Sci. USA 116, 7397–7402.

Pang, J., Liu, X., 2023. Evaluation of missing data imputation methods for human osteometric measurements. Am. J. Biol. Anthropol. 181, 666–676.

Perri, A.R., Koster, J.M., Otárola-Castillo, E., Burns, J.L., Cooper, C.G., 2019. Dietary variation among indigenous Nicaraguan horticulturalists and their dogs: an ethnoarchaeological application of the canine surrogacy approach. J. Anthropol. Archaeol. 55, 101066.

Pinhasi, R., Fort, J., Ammerman, A.J., 2005. Tracing the origin and spread of agriculture in Europe. PLoS Biol. 3, e410.

Porčić, M., Nikolić, M., 2016. The approximate Bayesian computation approach to reconstructing population dynamics and size from settlement data: demography of the Mesolithic-Neolithic transition at Lepenski Vir. Archaeol. Anthropol. Sci. 8, 169–186.

Price, M.H., Capriles, J.M., Hoggarth, J.A., Bocinsky, R.K., Ebert, C.E., Jones, J.H., 2021. End-to-end Bayesian analysis for summarizing sets of radiocarbon dates. J. Archaeol. Sci. 135, 105473.

Priß, D., Wainwright, J., Lawrence, D., Turnbull, L., Prell, C., Karittevlis, C., Ioannides, A. A., 2025. Filling the gaps—computational approaches to incomplete archaeological networks. J. Archaeol. Method Theor 32. https://doi.org/10.1007/s10816-024-09688-z.

Privat, K.L., O'connell, T.C., Richards, M.P., 2002. Stable isotope analysis of human and faunal remains from the Anglo-Saxon cemetery at berinsfield, Oxfordshire: dietary and social implications. J. Archaeol. Sci. 29, 779–790.

Ragno, R., 2024. Sheep and goats taxonomic abundance trends in 1st millennium CE southern Italy: multilevel bayesian modelling of NISP datasets. J. Archaeol. Sci. 171, 106068.

Riris, P., Silva, F., 2021. Resolution and the detection of cultural dispersals: development and application of spatiotemporal methods in Lowland South America. Humanit. Soc. Sci. Commun. 8, 1–13.

Riris, P., Silva, F., Crema, E.R., Palmisano, A., Robinson, E., Siegel, P.E., French, J.C., Jørgensen, E.K., Maezumi, S.Y., Solheim, S., Bates, J., Davies, B., Oh, Y., Ren, X., 2024. Frequent disturbances enhanced the resilience of past human populations. Nature 629, 837–842.

Rohrer, J.M., 2018. Thinking clearly about correlations and causation: graphical causal models for observational data. Adv. Methods Pract. Psychol. Sci. 1, 27–42.

Romanowska, I., Wren, C.D.A.C.S., 2021. Agent-Based Modeling for Archaeology: Simulating the Complexity of Societies. The Santa Fe Institute Press, Santa Fe.

Rosenstock, E., Ebert, J., Martin, R., Hicketier, A., Walter, P., Groß, M., 2019. Human stature in the Near East and Europe ca. 10,000–1000 BC: its spatiotemporal development in a Bayesian errors-in-variables model. Archaeol. Anthropol. Sci. 11, 5657–5690.

Rothman, K.J., Greenland, S., 2005. Causation and causal inference in epidemiology. Am. J. Publ. Health 95 (Suppl. 1), S144–S150.

Rubin, D.B., 1984. Bayesianly justifiable and relevant frequency calculations for the applied statistician. Ann. Stat. 12. https://doi.org/10.1214/aos/1176346785.

Rubin, D.B., 1976. Inference and missing data. Biometrika 63, 581.

Ryan-Despraz, J., Wissler, A., 2024. Imputation methods for mixed datasets in bioarchaeology. Archaeol. Anthropol. Sci. 16, 187.

Scott, E.M., Cook, G.T., Naysmith, P., 2007. Error and uncertainty in radiocarbon measurements. Radiocarbon 49, 427–440.

Shennan, S., 1997. Quantifying Archaeology. Edinburgh University Press.

Shipley, B., 2016. Cause and Correlation in Biology: a User's Guide to Path Analysis, Structural Equations and Causal Inference with R, second ed. Cambridge University Press, Cambridge, England.

Silva, F., 2020. A probabilistic framework and significance test for the analysis of structural orientations in skyscape archaeology. J. Archaeol. Sci. 118, 105138.

Silva, F., 2019. On measurement, uncertainty and maximum likelihood in skyscape archaeology. In: Visualising Skyscapes. Routledge, pp. 55–74.

Sisson, S.A., Fan, Y., Beaumont, M.A., 2018. Overview of Approximate Bayesian Computation arXiv [stat.CO].

Smith, J.A., Sandbrink, J.B., 2022. Biosecurity in an age of open science. PLoS Biol. 20, e3001600.

Smith, M.E., 2023. Making good arguments in archaeology. In: Gonzalez-Perez, C., Martin-Rodilla, P., Pereira-Fariña, M. (Eds.), Discourse and Argumentation in Archaeology: Conceptual and Computational Approaches. Springer International Publishing, Cham, pp. 37–54.

Sørensen, T.F., 2016. In praise of vagueness: uncertainty, ambiguity and archaeological methodology. J. Archaeol. Method Theor 23, 741–763.

Spaulding, A.C., 1953. Statistical techniques for the discovery of artifact types. Am. Antiq. 18, 305–313.

Stock, B.C., Jackson, A.L., Ward, E.J., Parnell, A.C., Phillips, D.L., Semmens, B.X., 2018. Analyzing mixing systems using a new generation of Bayesian tracer mixing models. PeerJ 6, e5096.

Surovell, T.A., Brantingham, P.J., 2007. A note on the use of temporal frequency distributions in studies of prehistoric demography. J. Archaeol. Sci. 34, 1868–1877.

Surovell, T.A., Finley, J.B., Smith, G.M., Brantingham, P.J., Kelly, R., 2009. Correcting temporal frequency distributions for taphonomic bias. J. Archaeol. Sci. 36, 1715–1724.

Szpak, P., Metcalfe, J.Z., Macdonald, R.A., 2017. Best practices for calibrating and reporting stable isotope measurements in archaeology. J. Archaeol. Sci. Rep 13, 609–616.

Thomas, D.H., 1978. The awful truth about statistics in archaeology. Am. Antiq. 43, 231–244.

Tsutaya, T., Yoneda, M., 2013. Quantitative reconstruction of weaning ages in archaeological human populations using bone collagen nitrogen isotope ratios and approximate Bayesian computation. PLoS One 8, e72327.

Vaiglova, P., 2025. How can we improve statistical training in archaeological science? J. Archaeol. Sci. 179, 106220.

van der Leeuw, S., 2016. Uncertainties. In: Uncertainty and Sensitivity Analysis in Archaeological Computational Modeling. Springer International Publishing, Cham, pp. 157–169.

Vanderplicht, J., Hogg, A., 2006. A note on reporting radiocarbon. Quat. Geochronol. 1, 237–240.

Van Pool, T., Leonard, R.D., 2011. Quantitative Analysis in Archaeology. John Wiley & Sons, Chichester.

Vieri, J., Crema, E.R., Uribe Villegas, M.A., Sáenz Samper, J., Martinón-Torres, M., 2025. Beyond baselines of performance: beta regression models of compositional variability in craft production studies. J. Archaeol. Sci. 173, 106106.

Wallace, I.J., Marsh, D. 'arcy, Otárola-Castillo, E., Billings, B.K., Mngomezulu, V., Grine, F.E., 2020. Secular decline in limb bone strength among South African Africans during the 19th and 20th centuries. Am. J. Phys. Anthropol. 172, 492–499.

Wallach, E., 2019. Inference from absence: the case of archaeology. Palgrave Commun 5, 1–10.

Wells, E.C., 2010. Sampling design and inferential bias in archaeological soil chemistry. J. Archaeol. Method Theor 17, 209–230.

Westreich, D., Greenland, S., 2013. The table 2 fallacy: presenting and interpreting confounder and modifier coefficients. Am. J. Epidemiol. 177, 292–298.

Whitehouse, H., François, P., Savage, P.E., Currie, T.E., Feeney, K.C., Cioni, E., Purcell, R., Ross, R.M., Larson, J., Baines, J., Ter Haar, B., Covey, A., Turchin, P., 2019. Complex societies precede moralizing gods throughout world history. Nature 568, 226–229.

Whittaker, J.C., Caulkins, D., Kamp, K.A., 1998. Evaluating consistency in typology and classification. J. Archaeol. Method Theor 5, 129–164.

Wissler, A., Blevins, K.E., Buikstra, J.E., 2022a. Missing data in bioarchaeology I: a review of the literature. Am. J. Biol. Anthropol. 179, 339–348.

Wissler, A., Blevins, K.E., Buikstra, J.E., 2022b. Missing data in bioarchaeology II: a test of ordinal and continuous data imputation. Am. J. Biol. Anthropol. 179, 349–364.

Wolfhagen, J.L., 2024. Estimating the ontogenetic age and sex composition of faunal assemblages with Bayesian multilevel mixture models. J. Archaeol. Method Theor 31, 507–556.

Youngblood, M., 2019. Conformity bias in the cultural transmission of music sampling traditions. R. Soc. Open Sci. 6, 191149.

Youngblood, M., Lahti, D.C., 2022. Content bias in the cultural evolution of house finch song. Anim. Behav. 185, 37–48.

Youngblood, M., Miton, H., Morin, O., 2023. Statistical signals of copying are robust to time- and space-averaging. Evol. Hum. Sci. 5, e10.