Statistical inference of prehistoric demography from frequency distributions of radiocarbon dates: a review and a guide for the perplexed

E. R. Crema*

* Department of Archaeology, University of Cambridge, Downing Street CB2 3DZ Cambridge, UK
Email: erc62@cam.ac.uk
ORCID No.: 0000-0001-6727-5138

Abstract

The last decade saw a rapid increase in the number of studies where time-frequency changes of radiocarbon dates have been used as a proxy for inferring past population dynamics. Although its universal and straightforward premise is appealing and undoubtedly offers some unique opportunities for research on long-term comparative demography, practical applications are far from trivial and riddled with issues pertaining to the very nature of the proxy under examination. Here I review the most common criticisms concerning the nature of radiocarbon time-frequency data as a demographic proxy, focusing on key statistical and inferential challenges. I then examine and compare recent methodological advances in the field by grouping them into three approaches: reconstructive, null-hypothesis significance testing, and model fitting. I will then conclude with some general recommendations for applying these techniques in archaeological and paleo-demographic research.

Keywords: Prehistoric Demography, Dates as Data, Statistical Inference, Radiocarbon Dates

Introduction

Population time series have a narrative appeal that has long been the envy of many archaeologists. Sister disciplines, such as economy and ecology, have developed methods, theories, and models that link individual-level processes to these macro-scale patterns and have inspired generations of archaeologists to find ways to borrow and extend these concepts to the study of the human past. The opportunity to generate something that visually resembles population time series is a source of major temptation — all those ideas and concepts can finally be applied to understand the archaeological record. Thus, it comes as no surprise that the so-called *dates as data* (hereafter DAD) approach (Rick 1987), which relies on the assumption that the changing frequency of radiocarbon dates related to anthropic events is a reliable proxy of relative past population change, is a low-hanging fruit that has been harvested extensively in the last decade.

Inferring population trajectories from time-frequency data is hardly a novel concept and certainly not limited to radiocarbon dates. Archaeologists have long been and are still counting different *things* as a proxy of population size, ranging from classic examples such as sites, dwellings, or artefacts (Drennan et al. 2015 for a review) to less common applications like faecal stanols (White et al. 2018). What makes DAD different, and in many cases controversial, is the unspecified nature of the *thing* that is being counted. Sites, dwellings, potsherds, and faecal stanols represent unique categories of artefacts that can be more or less directly related to specific behavioural processes. On its own, radiocarbon dates are just numerical attributes of virtually anything carbon-based and relate to a highly diverse range of anthropic and non-anthropic processes. Population inference based on radiocarbon dates does not necessarily have to subscribe to the DAD assumption, and time frequencies can relate to specific types of events (e.g. use of residential features, cf. Oh et al. 2017). More broadly, radiocarbon frequency data have also been used to examine cultural phenomena such as changes in burial or subsistence practices (e.g. Stevens and Fuller 2012, Glesson and McLaughlin 2021) and hence their analyses are not restricted to the reconstruction of past population dynamics either. These examples, where events associated with the radiocarbon record are well defined, should not be referred to as DAD. The main appeal and the primary issue with Rick's approach stem from the tactical decision of prioritising larger sample sizes at the cost of being vaguer on the nature of the dates to be included in the analysis.

There is, however, a separate and additional layer of complexity, issues, and challenges dictated by the statistical nature of the method proposed. Some of these are not specifically limited to radiocarbon dates and are relevant to other attempts in inferring population changes from archaeological frequency data (see Brown 2015 for discussion); namely, the: 1) non-random and systematic nature of chronological uncertainty; 2) the problem of *sampling error*; and 3) and the substantially wide range of possible population curves that we are aiming to reconstruct. The intersection of these three broader issues makes any frequency analyses of radiocarbon dates challenging, even when issues about the nature of the proxy or the definition of the events associated with each date are addressed. More importantly, there are no readily available, off-the-shelf solutions to many of these analytical problems. Consequently, the last few years saw the proposal of a substantial wide range of new statistical approaches developed in prehistoric population studies.

This paper aims to review and compare the current range of statistical methods designed to analyse time frequencies of radiocarbon dates. Over the last few years, several review papers

have examined different aspects of radiocarbon based population inference, including the problematic nature of the proxy (Attenbrow and Hiscock 2015); the misleading effects of the calibration process (Williams 2012, Weninger et al. 2015); the importance of growth rates (Brown 2017) as well as their comparability to ethnographic scales (Tallavara and Jørgensen 2021); and the critical issue of radiocarbon sampling processing (Becerra-Valdivia et al. 2020). A systematic review of more recent methodological solutions does not exist, as most discussions on the statistical nature of the problem are either limited to small sections of papers arguing in favour of particular solutions (see, for example, Brown 2015, Crema et al. 2017, Bronk-Ramsay 2017, Timpson et al. 2021, Carleton 2021), or broader criticisms of particular methodology such as the summed probability distribution of calibrated radiocarbon dates (hereafter SPD, Carleton and Gourcutt 2021). The substantially wide range of statistical options available and the idiosyncrasies of contextual issues have made the whole research area harder to navigate. As a result, unwarranted criticisms are often raised without a clear understanding of what a particular method entails, whilst simultaneously, there is an increased risk of misuses, abuses, and misinterpretations of these novel solutions. The objective of this paper is also to focus the spotlight on neglected key details that are often hidden behind equations or lines of code or implicit in the description of particular techniques. In most cases, these details have no impact in qualitative terms, but there are circumstances where conclusions can be drastically different.

From *dates as data* to Summed Probability Distributions

Rick's seminal paper first introduced the core assumption that '[a]ll things equal, more occupation produced more carbon dates' (Rick 1987, 56), immediately acknowledging in the following sentence that such an equation will be affected by a variety of intervening factors, most notably *creation*, *preservation*, and *investigation* biases (*ibid*. Fig.1). The original approach simply consisted in creating histograms of uncalibrated ^{14}C ages. Still, it was already coupled with more advanced techniques, such as bootstrap confidence intervals to consider potential spurious effects emerging from sampling error (*ibid*. Fig.4). The approach had some discrete success already in the early 90s when several authors have switched from histograms of uncalibrated ^{14}C ages to curves generated using calibrated dates (e.g., Ames 1991, Dye and Komori 1992, Erlandson et al. 1992, Chatters 1995). Some of these early applications have also led to the development of new statistical techniques, such as randomisation tests[1] (Dye 1995), or even attempts to combine historical census data and inferred growth rates to retrodict absolute (rather than relative) population sizes for the pre-census era (Dye and Komori 1992). The transition from the summation of uncalibrated to *calibrated* ^{14}C ages became problematic once the calibration process no longer made it possible to describe calibrated dates using symmetric errors. In response to an early work by Hounsely et al. (1997), who summed uncalibrated dates using Gaussian distributions and a moving sum, Blockley and colleagues (2000) stressed that uncalibrated dates would provide unreliable results as they are based on a different, non-linear timescale. They then argued that "[o]nce dates have been calibrated they can no longer be expressed as a point date with a Gaussian error because the probability distribution of the date is a function of the shape of the calibration curve [...] Because of this, a moving sum which gives no weight to the actual probability distributions of dates is unlikely to be a good assessment of their true distribution. It is more appropriate to look at the *summed probability distributions of the calibrated dates* [...]" (emphasis added). As far as I am aware, this was one of the earliest applications of what

1 While preparing this manuscript, I came across a paper by Tom Dye. He was the first to introduce randomisation tests to compare curves generated from the summation of calibrated radiocarbon dates. In 2016, I have, together with my colleagues, effectively reinvented the wheel by introducing a similar technique to compare regional demographics in prehistoric Japan (Crema et al. 2016).

is now undoubtedly the most common form of radiocarbon frequency analyses, often now simply referred to as SPD.

The first significant criticisms against SPDs were raised a few years later by Blackwell and Buck (2003) in the context of reviewing previous works on the Late Glacial human occupation in north-western Europe (including both Hounsely et al. 1997 and Blockley et al. 2000) and advocating for a model-based Bayesian solution as a more robust alternative. Their review stress two core issues: 1) the problematic nature of summing probabilities; and 2) the fact that "since the calibrated dates being 'summed' do not relate to the same event, it is not clear what interpretation can be placed on the probabilities produced by this method" (*ibid*, page 233). While Blackwell and Buck do not provide much detail for the first problem, it is reasonable to assume that this relates to the mathematical issue of how summed probabilities are no longer probabilities, and whilst representing in some way the density distribution of the phenomena of interest, they cannot be straightforwardly interpreted (see Carleton and Gourcutt 2021 for a recent exhaustive review on this issue) as they mask the uncertainties inherited from individual dates. For example, consider a scenario where two time intervals, t_1 and t_2, are both associated with a summed probability of 10. Now suppose that t_1 contains ten radiocarbon dates, each with a probability of 1, while t_2 has 100 radiocarbon dates, each with a probability of 0.1 for that interval. In other words, we are sure that ten events are associated with t_1, while we have much more uncertainty for t_2. Summed probability cannot distinguish the two and simply conveys a message that there was no change in the number of events from t_1 and t_2 without providing a measure of uncertainty on such a claim. In this particular case, the probability that t_2 has exactly ten events is only 0.13, with a probability of increase from t_1 to t_2 equal to 0.41 and a probability of decrease equal to 0.45[2].

The second issue raised by Blackwell and Buck concerns the core assumption of *dates as data*, i.e., what is being counted are simply *dates,* and the events they are associated with are ambiguously defined (e.g. 'anthropic'), encompassing a wide range of behavioural processes. Rick's gambit hinges on the assumption that the aggregate frequency of radiocarbon dates associated with different anthropic events correlates with population density, retaining a reliable signal by evening out its underlying heterogeneity. A relatively large number of papers have discussed how this assumption can be problematic (Attenbrow and Hiscock 2015, Torfing 2015, Becerra-Valdivia et al. 2020, Ward and Larcombe 2021). While this is unquestionably an important issue, I will not add much more to the debate for two reasons. Firstly, the problem is context-dependent — demonstrating that the assumption does or does not hold for a particular dataset does not allow its conclusion to be generalised to all DAD applications. Secondly, the problem arises prominently if events associated with the sample dates are not clearly identified. In other words, if one decides to limit their dataset to radiocarbon dates associated with particular types of events (e.g. the constructions of dwellings), much of the issue is reduced to the extent by which the correlation between the frequency of such events and the population under investigation is stationary over time (and space). Of course, this does not necessarily solve all interpretative problems. Still, it is worth noting that time-frequency analyses of radiocarbon dates represent a wider class of analyses, models, and issues than DAD.

2 These probabilities can be computed using the binomial probability mass function.

The curse of eyeballing

The issues discussed in the previous section are just a fraction of a wider range of problems associated with the direct interpretation of SPDs discussed in the literature. While readers concerned with these problems should consult more detailed discussions for each, it is worth briefly revisiting some of the key matters raised, namely: 1) sampling error; 2) heterogeneity in sampling intensity; 3) spatial averaging and non-stationarity; 4) taphonomic loss, and 5) systematic measurement errors associated with the calibration process.

Sampling Error

A trivial (but somewhat surprisingly too often disregarded) aspect of time-frequencies of radiocarbon dates (or any other count-based population proxy) is the notion that the observed data are just *samples* and not the statistical population. A simple way to conceptualise this is to consider the observed sample of dates as random draws from a probability distribution spanning the time window of interest and characterised by an unknown shape that we aim to recover. This effectively formalises the assumption of any frequency-based proxy — we expect to find *more* 'things' (e.g., sites, artefacts, radiocarbon dates) during intervals where there are *more* people; if we have twice as many people for a given time interval, we should expect twice as many 'things' we are counting. In practice, however, this relationship is conditioned by the available number of dates, and observed data can deviate from this expectation. In other words, even if there is a perfect correlation between human population size and the frequency of radiocarbon dates, there will always be some deviations arising from sampling error and observed peaks and troughs might not be a genuine signal of population change. As mentioned earlier, the problem was already raised in Rick's original work and has since then been tackled in a variety of ways (e.g. Michczyńska and Pazdur 2004, Kelly et al. 2013, Shennan et al. 2013, Manning and Timpson 2014, Brown 2015, Dye 2016, Bronk Ramsey 2017). Larger sample sizes can, of course, minimise the problem of sampling error, and as such, it is tempting to think whether there is a threshold above which the problem can be safely ignored. A widely cited work by Williams (2012) has for example, provided a guideline figure of 500 dates, following previous simulation-based analyses by Michczyńska and Pazdur (2004) and by Geyh (1980). While a clear answer to the question 'how many dates do I need for my SPD?' might sound reassuring, the reality is that this ultimately depends on the scale, the granularity, and the magnitude of the specific fluctuations we wish to identify (see Hinz 2020 for a simulation-based study on this problem). To a large extent, this is akin to the issue of statistical power in null significance hypothesis testing (NSHT); sample size is only one side of the coin, and its required value depends on the effect size we wish to determine. Large trends can be detected from smaller sample sizes while identifying smaller fluctuations requires more data. The problem is exacerbated by the fact that we have much less clue about the shape of the target population compared to other kinds of data. For example, if we were to examine a small sample of femur lengths from a particular cemetery assemblage, we would expect, *a priori*, a normal distribution following the central limit theorem — if we plot a histogram and observe a small deviation from a bell-curve we would be inclined to dismiss this as the result of sampling error. The frequency distribution of radiocarbon dates has fewer and much less formalised general principles that can help us be sceptical about the peaks and troughs we observe. Aside from extreme fluctuations, we would regard many of the patterns we observe as plausible evidence of population change. In other words, we do not have a strong prior on the expected shape of the SPD, and having an epistemic stance prone to over-interpretation does not help.

Heterogeneity in Sampling Intensity

Adopting formal statistical inference (see next section) can address the problem of sampling error. However, this is ensured only if the two fundamental assumptions of statistical samples - randomness and independence - are met. Radiocarbon dates are clearly not randomly sampled from a population of possible dateable artefacts. In most cases (but see Porčić et al. 2021 for an exception), samples for demographic inference are based on the re-use of 14C dates collected for a wide range of purposes using various sampling strategies. The question is whether, with a sufficiently diverse set of sampling strategies and designs underlying a given dataset, we can treat the sample *as if* it were random. The answer is, once again, context-dependent, but there are a few typical cases where such an assumption does not hold. The most notable one is that the likelihood of employing radiocarbon dating declines when investigating historical periods where more accurate, precise, and cheaper dating methods become available. It follows that all radiocarbon-based time-frequency data suffer from an edge effect approaching the present day, with a magnitude and timing that vary geographically and limit opportunities for cross-regional studies for more recent periods.

Systematic temporal variations in sampling intensity are harder to detect when they are likely to produce biases that do not contradict our expectations as bluntly as the case of the declining density towards the present day. For example, one could postulate that an increased interest in dating more accurately the earliest evidence of Neolithisation might promote a higher sampling intensity and consequently lead to a higher density of radiocarbon dates during the early stages of the Neolithic period. The problem here is that we also expect an increase in the population size during this period, and as such, we would hardly interpret a higher density of radiocarbon dates during this interval as an anomaly or the consequence of a research bias. Heterogeneous sampling intensity across time is perhaps the most concerning and simultaneously less understood bias that might affect the DAD approach. One possible way to mitigate its impact is to include statistical variables aimed to control the potential impact of the original purpose of dating, e.g., by discerning dates from specific research projects to those obtained in rescue excavations. While no attempts have been made in this direction yet, statistical analyses of different recovery practices do show specific signatures (Vander Linden 2019) and might provide a baseline for accounting for these kinds of biases.

The mixture of different objectives and dating practices is particularly evident when examining inter-site variations in sampling intensity. For example, in the EUROEVOL database (Manning et al. 2014), the largest number of dates associated with an individual site is 184, whilst more than half of the sites (2,138 out of 4,213) contained only a single date. Several solutions have been proposed to tackle this problem. For example, dates that are known to be referring to the exact same event can be combined following Ward and Wilson's method (1978; see for example Ahn and Hwang 2015). A similar procedure often referred to as 'binning' (see Timpson et al. 2014), consists of generating a 'local' SPD by summing the calibrated probability of dates from the same site that are 'close' in time and normalising to sum to unity the area of the resulting curve. In both cases, the net result is to treat sets of multiple dates as one and effectively compensate for the unevenness in sampling intensity. There are, however, different implications between the two approaches. In the first case, the aggregation process does not alter the nature of what is being counted as it relies on the notion that sets of dates refer to the same event. Thus, for example, if dates are aggregated based on the construction of residential units (our target event), the resulting frequency data would still be a proxy of changes in the number of dwellings over time. The situation is slightly different in the case of the 'binning' approach. Here the aggregation 'ensures that each *site-phase* is equally weighted when generating the SPD' (Timpson et al. 2021,

emphasis added), which implies that effectively we are defining the target as loosely defined 'site occupation' counts. The problem becomes even more complex as the 'binning' approach requires some temporal threshold for aggregating dates that are 'close' in time. Modifying such a threshold could yield rather different results, and while one can carry out sensitivity analyses, the nature of what is being counted remains hostage to the value assigned to such parameter. Shifting the interpretation of the temporal frequencies of radiocarbon dates from 'population size' to 'number of occupied settlements' can help, but at the same time, this introduces interpretative consequences. Empirical estimates of growth rates obtained can no longer be assumed to be directly emerging from demographic events (i.e., birth, death, and migration) alone but rather as a joint outcome of these processes with episodes of settlement fission, fusion, and extinction. Shifts between nucleated and dispersed settlement patterns, changes in the duration of settlement occupation, or variations in intra- and inter-annual residential mobility patterns are just some examples of processes that can lead to signals *without* an actual change in the underlying human population (cf Bevan and Crema 2021). This is a problem of interpretation, and while it does not on its own jeopardise the DAD approach, it further emphasises the issues of comparability between growth rates estimated from archaeological data to those observed in ethnographic and historical contexts (see Tallavaara and Jørgensen 2021), or even how differences between different archaeological population proxies should be interpreted (see Palmisano et al. 2017, Crema and Kobayashi 2020, Seidensticker et al. 2021)

Spatial Averaging and Non-stationarity

The ubiquity of radiocarbon data and the increasing availability of larger databases (e.g., Manning et al. 2016, Chaput and Gajewski 2016, Lucarini et al. 2020, Martínez-Grau et al. 2021, Bird et al. 2022) has pushed many to attempt reconstructing prehistoric population dynamics for larger windows of analyses, often at continental scales (Williams 2012, Shennan et al. 2013, Wang et al. 2014)

Summarising putative population dynamics of a vast geographic area with a single time series can undoubtedly be misleading, as it implicitly assumes that all sub-regions had similar demographic trajectories. The trade-off is between selecting a smaller window of analyses that accounts for spatial variation but is impacted by higher sampling error or opting for a wider region that benefits from a larger sample size but yields a 'space-averaged' estimate (Porčić et al. 2021) that might not be representative of any of its sub-regions. The problem is further exacerbated by the fact that larger study areas are likely to be characterised by variations in sampling strategies and intensity, as different administrative and geopolitical units are often associated with substantial variation in wealth, sample design, and research interests (Crema 2020).

The use of spatial analyses that explicitly explores regional variation in demographic trajectories (Timpson and Manning 2014, Chaput et al. 2015, Crema et al. 2017, Riris and Arroyo-Kalin 2019) can offer far more informative insights for larger regions than a single timer-series. However, as for frequency time-series, these spatio-temporal density maps cannot be based exclusively on visual assessment and needs explicitly account for variations in sampling intensity (e.g., using relative risk surfaces; see Chaput et al. 2015, Bevan et al. 2017) as well as the delicate balance between spatial resolution and sampling error (e.g., by using spatial permutation tests Crema et al. 2017, Riris and Arroyo-Kalin 2018)

Taphonomic loss

Taphonomic loss, and other post-depositional processes, are another key factor that can bias the raw and direct interpretation of the radiocarbon record and other types of time-frequency data. As for many of the other biases discussed above, the issue was already raised in Rick's seminal paper, which, amongst other things, highlights the implication of older dates being less likely to survive and included in the sample. A model-based assessment of the potential magnitude of taphonomic loss has been explored by Surovell and Brantingham (2007), who showed how under extreme conditions, an exponentially declining population could even yield an exponential growing frequency curve. Adjusting frequency data for taphonomic loss is straightforward but requires a loss function derived from independent estimates. Surovell and colleagues have (Surovell et al. 2009, see also Bluhm and Surovell, 2018 for an updated version) used radiocarbon ages from volcanic deposits to empirically estimate the impact of taphonomic loss. Their analyses revealed that the rate of taphonomic loss is not constant, but declines as the age of the site grow and propose a global 'correction formula' that accounts for this factor for time-frequency data between 40,000 and 1,000 cal BP. The implication of this correction can vary between datasets and is generally expected to have a greater impact when dealing with multi-millennial scales. Still, several studies have also reported negligible effects (see for example Zahid et al. 2016, Tremayne and Winterhalder 2017, Broughton and Weitzel 2018, Fernández-López de Pablo et al. 2019).

Calibration Effects

The uncertainty associated with radiocarbon dates is a combination of sample-specific measurement errors and the systematic effect of the information loss resulting from the calibration process. The random nature of the former makes it a comparatively negligible factor for most objectives, with limitations primarily concerning the analytical resolution. With a sufficiently large sample size, the impact of these errors can, in most cases, be considered negligible. The systematic nature of the latter is far more problematic as it can lead to artificial patterns in the time-frequency data — with all other things being equal, 14C dates within calibration 'plateaus' will tend to produce wider and flat calibrated probability distributions. In contrast, samples located within steeper portions of the curve will tend to have narrower and more 'spiky' distributions (but see Brown 2015). In this case, increasing the sample size does not help — the sum of flat probability distributions with similar ranges will, unsurprisingly, be a flat probability distribution. The cumulative consequence of this effect is that some of the fluctuations observed in empirical SPDs are just the results of these calibration effects. This is a well-known problem that has been pointed out repeatedly in the literature (Guilderson et al., 2005 Williams 2012, Brown 2015, Wenninger et al., 2015, Crema and Bevan, 2021).

It is worth noting that the problem is not unique to radiocarbon dates and applies to any dating method where events closer in time have similar systematic information loss. Perhaps the most common example is the use of archaeological periodisations and relative chronologies, and its implications become tangible when attempts are made to quantify their uncertainty and convert assignments to particular periods or phases into absolute calendar dates. Several approaches have been proposed in the literature, starting from the application of aoristic analysis (Johnson 2004, Crema 2012) to the use of more complex probability models (Baxter and Cool 2016, Collins-Elliot 2019, Crema and Kobayashi 2020) to convert a given 'time-span' of the possible existence of an event into a probability distribution. The issue, in this case, is that the extent of such temporal intervals is in practice informed by the presence of some diagnostic features which allow the specialist to assign a particular object

into a phase (e.g., 'Early Bronze Age I'). Thus, two events that are separated in time, but have similar diagnostic features, will be assigned to the same "time span of existence", and ultimately have identical probability distributions. It follows that summing these probabilities (e.g. using 'aoristic sums') will yield time-series with spurious artefacts similar to those observed in SPDs (see Bevan and Crema 2021 for discussion).

Calibration effects have been tackled mainly by applying some smoothing techniques to remove indiscriminately any short-term fluctuations in the SPDs. These can be as simple as calculating the average summed probability over a sliding window (e.g., Shennan et al. 2013, Kelly et al. 2013) or more complex solutions involving the joint use of Monte-Carlo simulations and Kernel Density Estimates (e.g., Brown 2017). These and other solutions (e.g., Wenninger et al. 2015) can help deter over-interpretations of radiocarbon frequency data, particularly for shorter temporal scales (<500 years) where the impact of these systematic errors is particularly pronounced. However, it is worth noting that many of these methods are effectively designed to "mask" the effect of calibration for visualisation purposes and do not address the problem directly and systematically.

Statistical inference

The brief survey of potential biases affecting radiocarbon time-frequency is a reminder of how visual inspections of SPDs should be carried out with extreme caution. Any insights obtained from visual assessments should be appropriately examined to formally discern whether they pertain to processes of interest or are mere statistical artefacts. While this principle generally applies to data visualisations, the lurking temptation of making post-hoc narratives from SPD plots appears to be particularly common despite continuous reminders and warnings in the literature to consider potential confounding factors.

The confidence that SPDs can be read as a *direct* signal of fluctuations in radiocarbon density (and conversely in population density) has led many to take a further step and carry out statistical analyses *directly* using the temporal sequence of summed probability values in SPDs. Examples range from simple correlations between SPD curves and other time-series such as paleoenvironmental data (Palmisano et al. 2021) or other population proxies (Crema et al. 2020) to more sophisticated analyses, including the use of Granger causality analyses to explore lagged responses to climatic events (Kelly et al. 2013), attempts to identify early warning signals of collapse (Downey et al. 2016), or uses ecological population models (Freeman et al. 2021) with externally induced, time-varying carrying capacities (Lima et al. 2020). The level of sophistication achieved by some of these studies is often very high and undoubtedly offers a glimpse of the kind of exciting questions that we could answer. Yet, fundamental concerns regarding sampling error or calibration effects are often ignored or just mildly acknowledged without a formal exploration of what their impact would be.

The extent to which inferences based on direct statistical assessments of SPDs are biased will inevitably depend on the specific context, but the general expectation is that this is a function of sample size, absolute time-interval, and the temporal granularity of the process under investigation. When sample sizes and the chronological granularity of the analyses are sufficiently large, the impact of sampling and calibration is likely negligible compared to the signal we aim to detect. However, there is no simple way to determine when this is the case. How many radiocarbon dates do we need to stop being concerned about sampling error? What is the appropriate temporal scale of analyses so that the impact of calibration can be safely ignored? As it is always in these cases, the answer is an unworkable and unsatisfying

'it depends'. As noted by Price et al. (2021), even with an infinitely large number of radiocarbon dates, an SPD would not be able to recover the shape of the underlying population as a result of the summation of the probabilities and the systematic impact of calibration.

There are situations where ignoring these issues can lead to strikingly different outcomes. For example, Lima et al. (2020) have recently constructed an SPD for the Pacific Island of Rapa Nui and fitted different logistic growth models. They utilised information criteria to demonstrate that the highest support was found in a model where the carrying capacity was a function of environmental covariates, which they used as an argument in support of the so-called *ecocide* hypothesis. A follow-up study by Di Napoli et al. (2021) employing Approximate Bayesian Computation (see below for details), which accounts for sampling error and calibration effects, has shown no support for such a model and instead indicated that, with the available evidence at hand, there was no way to discern between the competing models.

However, the direct use of SPD values for statistical analyses does not represent the entirety of inferential approaches dedicated to population studies based on time frequencies of radiocarbon dates. In less than a decade, a significant number of novel methods that account for many of the issues discussed in the previous section have been proposed in the archaeological literature. They all share a fundamental dissatisfaction with approaches based on the *direct* interpretation of SPDs and offer solutions tailored to specific inferential needs (see below and Table 1 for a summary). Despite some fundamental differences, these techniques can be broadly classified into three groups based on their primary objective: 1) *reconstructive* approaches, 2) *null-hypothesis significance testing (NHST)* approaches, and 3) *model-fitting* approaches. As for any attempts in imposing sharp categorical boundaries, one should be critically aware that many of the methods presented below do share conceptual roots, and a combination of techniques from different approaches can well coexist in the same study.

Reconstructive approaches

The section above has repeatedly highlighted that a visual inspection of SPDs is not warranted and may lead to biased interpretations in some situations. Yet data visualisations can be a powerful tool to highlight information that cannot be sufficiently portrayed by numbers alone (Anscombe 1973). Thus, it does not come as a surprise that many have attempted to tackle this difficult trade-off by implementing a visualisation technique that can simultaneously correct for the impact of the calibration process whilst acknowledging the potential impact of sampling error by displaying an envelope surrounding observed SPD values.

A few different approaches have been proposed to achieve this objective (see Table 1 and Fig. 1), with the earliest application dating back to the already mentioned bootstrap confidence interval employed by Rick (1987, fig. 4). Since then, other authors have taken a similar approach (e.g., Timpson and Manning 2014), sometimes in conjunction with more sophisticated procedures. For example, McLaughlin (2019) advocates a solution based on a combination of bootstrapping and kernel density estimates. Given a collection of radiocarbon dates, the approach consists of 1) randomly selecting (with replacement) a subset of the sample; 2) calibrating the sampled dates; 3) sampling a calendar date from each calibrated probability distribution, and 4) running a univariate kernel density estimate (KDE). The process is repeated multiple times so that an ensemble of KDEs is obtained, combined, and

visualised as an envelope (Fig. 1, first row; see also Brown 2017 for a similar approach but without the bootstrapping step). Such bootstrapped composite KDE (cKDE) addresses the issue of sampling error (step 1), chronological uncertainty (step 3) and the problem of calibration artefacts (KDE smoothing in step 4). The choice of bandwidth size and the shape of the kernel can have a significant impact on the final product, with the resulting curve being either under or over-smoothed. McLaughlin suggests a comparatively small bandwidth (e.g., 30 years) for most applications to capture sudden changes in density, but it is an open question whether this size can avoid all instances of artificial calibration peaks often observed in SPDs. While there are a relatively large number of algorithms designed to find optimal bandwidth sizes based on the observed data (Heidenreich et al. 2013), there is no clear consensus on which one should be preferred, nor a systematic exploration of which methods are better suited for demographic inference. Finally, KDEs are typically affected by an edge effect, with a decline in density at the start and the end of the window of analysis. Edge correction formulas do exist, but their application becomes problematic given the nature of the resampled data, and the most straightforward approach seems to be the selection of a wider data window and a narrower visualisation window.

The problem of bandwidth size selection can be solved by treating this as a parameter to be estimated using Bayesian inference. This solution was developed by Bronk-Ramsey (2017) and is implemented in the widely used calibration and Bayesian analyses software *OxCal* (see Fig. 1: second row). The approach consists of using a uniform prior for the bandwidth size h with an upper limit based on Silverman's rule (1986), which provides a criterion for identifying h when the underlying distribution is Gaussian. Bronk-Ramsey considers this as an upper threshold that would over-smooth multimodal distributions. The predictive likelihood used to estimate h is instead based on the product of likelihoods of each date as modelled by the KDE based on the remaining data, excluding the focal date. The model can be fitted alongside other distribution models in OxCal (e.g., uniform, Gaussian, exponential, etc.) that will act as a prior and can modify the shape of the kernel for each date. Alternatively, an extension of this approach (called *KDE_Model* in OxCal) can be adopted where the prior for each observation point is effectively the KDE distribution of all the other radiocarbon dates.

While the KDE approach proposed by Bronk-Ramsey has both elements of frequentist and Bayesian inference, a full non-parametric Bayesian approach is also possible via the finite Gaussian mixture model (Fig 1: third row). This is a flexible method that is now widely used in many fields (see for example in isotopic studies Fernandes et al. 2014) and the *Bchron* (Haslett and Parnell 2008) and the *baydem* (Price et al. 2021) R packages offer functionalities for its application for radiocarbon analyses, albeit with some minor differences in their implementation. The core idea of a finite Gaussian mixture is to conceive the observed data as the aggregation of a finite number of Gaussian distributions, each with its own mean and standard deviation. The inferential process consists of determining the number of mixture components (i.e., Gaussian distributions), their associated parameters (i.e. mean and standard deviation), and their relative contributions (i.e. expected proportion of the data), which provides a flexible range of probability distribution shapes. In contrast to other applications (e.g., isotope-based diet reconstructions), the objective here is not the recovery of particular parameters but the overall shape of the probability distribution, which effectively portrays how the density of radiocarbon dates changed over time whilst accounting for sampling error and calibration effect. Price et al. (2021) have recently developed this technique specifically for the use of demographic archaeology by stressing the importance of the direct computation of the likelihood (see also below). They provide a Bayesian workflow and an associated R

package to facilitate its application (*baydem*), allowing users to assign specific priors or to estimate the optimal number of mixture components. They illustrate their technique by examining the radiocarbon record of the Maya city of Tikal, showing how their approach is consistent with previous studies based on other lines of evidence and proxies, whilst providing a more precise estimate of the timing of key demographic events.

The three approaches discussed above provide more robust alternatives to SPD for visualising the radiocarbon density record. One of the most appealing aspects shared by all solutions is that, in contrast to other methods described below, some of them require a relatively smaller number of assumptions by the end-user. OxCal's KDE can be fully automated, cKDE requires only the number of bootstrap iterations and the kernel bandwidth size. Bayesian finite Gaussian mixture models do, however, require additional user-defined settings, including hyperparameters and the number of mixture components. The latter is a key parameter as it defines the complexity of the resulting shape of the density distribution, but users can specify multiple values and carry out model selection via Pareto smoothed importance sampling (PSIS) to determine the optimal number whilst avoiding overfitting. There is, however, a substantial variation in terms of computational costs. cKDE with bootstrapping is a relatively fast method that will take just a few minutes even when the sample size is relatively large; *baydem*'s Bayesian finite Gaussian Mixture Model would require a much longer processing time, especially when dealing with larger sample sizes and the range of mixture components to be explored is high. OxCal's KDE comes with the highest computational cost, with runtimes ranging from several hours to a few days when the sample size is above 1000 dates. Despite these differences in computational costs, the difference in the output (particularly about the "true" population) can be negligible in many situations (Fig 1, see also figure 2 in Price et al. 2021), particularly when sample sizes are large.

In contrast to the other methods detailed below, these reconstructive approaches can be seen as the go-to solution for any preliminary assessment of the available data. These approaches are particularly appealing because they do not require the user to assume a priori a specific shape of the underlying density distribution. However, there are two things to consider. The first relates to the unavoidable weakness of all three approaches when dealing with smaller sample sizes (see Fig.1, third column). Confidence envelopes are larger in these cases, but they might still fail to include the true underlying probability distribution. Unfortunately, because of the very nature of these models, there is no way to determine an optimal minimum sample size as this would depend on the scale and magnitude of the signals one is hoping to reconstruct. The second issue stems from the fact that these tools can be abused as inductive inference engines. The confidence that visual outputs produced by these methods are more reliable than SPDs can easily entice scholars to develop post-hoc explanations without formal and direct testing.

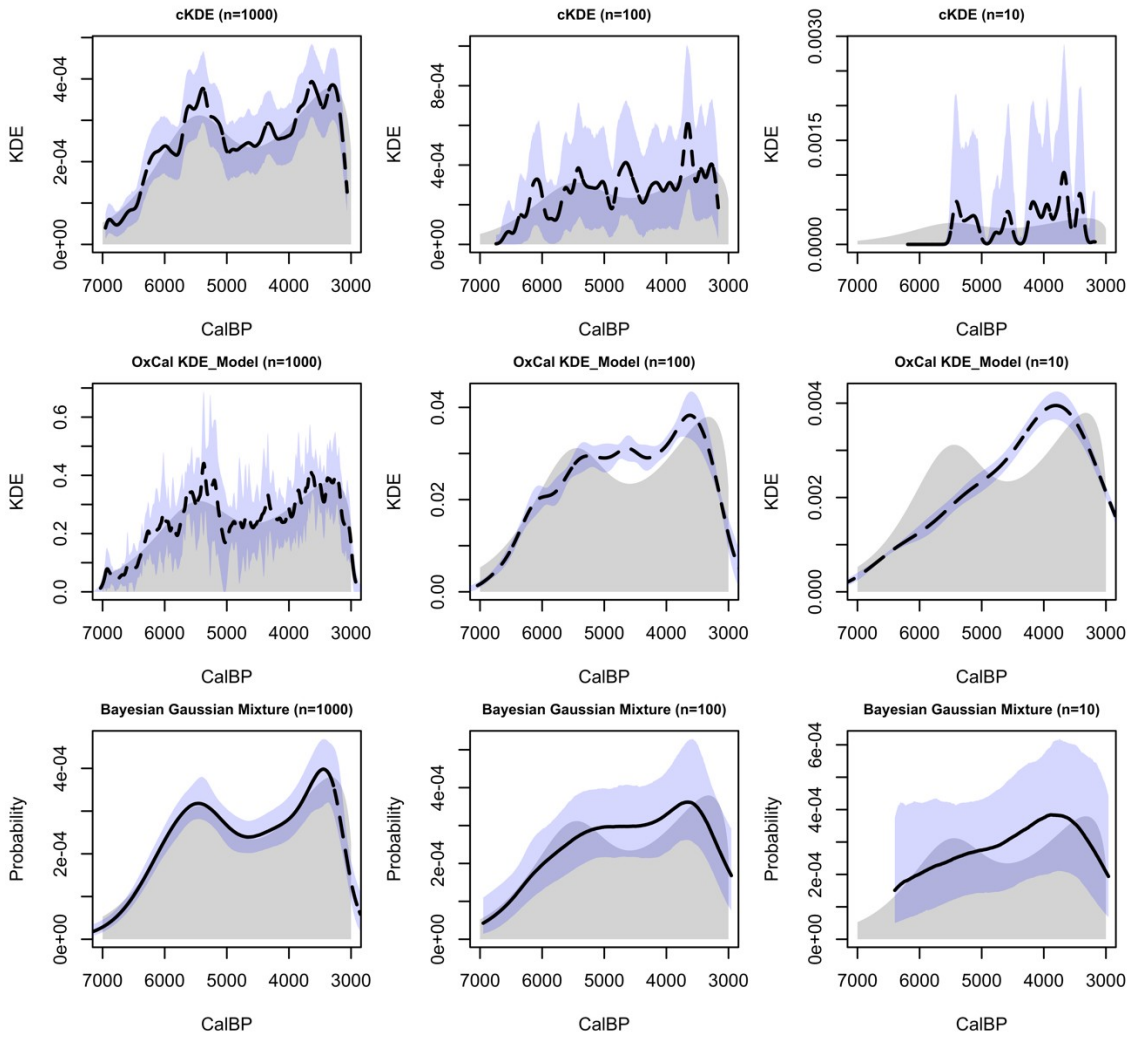

Figure 1. *Comparison of reconstructive approaches to radiocarbon frequency data on small (n=10), medium (n=100), and large (n=1000) datasets using bootstrapped Composite Kernel Density Estimate, OxCal's Model_KDE and baydem's finite Gaussian Mixture model. The grey area represents the shape of the underlying probability (identical for the three sets) from which radiocarbon dates were sampled from. R scripts required for generating the figures are available at https://github.com/ercrema/c14demoreview and archived on zenodo (https://doi.org/10.5281/zenodo.6421345).*

Null-Hypothesis Significance Testing (NHST) approaches

Approaches in this category are designed to address the limitation of reconstructive methods by formally examining *specific* hypotheses. For example, one might be interested in determining whether observed time frequencies of radiocarbon dates conform to or deviate from what we should expect from an exponential population growth with a particular rate or whether two regions have experienced similar population trajectories during a specific time window. These examples are well suited for applying a Null-Hypothesis Significance Testing (NHST) framework.

The number of case studies employing NHST for examining radiocarbon time-frequency data has grown substantially since the publication of the seminal paper by Shennan and colleagues (2013), who first introduced a Monte-Carlo simulation approach that underpins most of the current applications. A comprehensive review of these approaches and an introduction to a dedicated R package that facilitates their applications is provided elsewhere (Crema and

Bevan 2021), but it is worth highlighting here the core idea behind these methods and more importantly, their limitations in practical applications.

The Monte-Carlo simulation approach introduced by Shennan et al. (2013) consists of comparing the observed SPD against a *distribution* of SPDs that one should expect to obtain given a particular null model. The intuition here is that given a growth model and a sample size of radiocarbon dates, one can iteratively generate an ensemble of SPDs and determine whether the observed SPD can be distinguished from those or not. In practical terms, such a null model is conceptualised as a sequence of probabilities values associated with each calendar year, e.g. P(t=2500 BP) = 0.001, P(t=2499 BP) = 0.002, P(t=2498 BP) = 0.003, etc. This effectively formalises the simple notion that if a particular year is assumed to have twice the population size of another, we would assume that the number of expected dates (hence the associated probabilities) would be two times larger. This discrete probability distribution is used to simulate *n* dates, with *n* equivalent to the observed sample size. The resulting set of calendar dates is then converted into ^{14}C age by 'back-calibration', and a measurement error, sampled with replacement from the observed data, is randomly assigned to each. This workflow generates *n* radiocarbon dates that we should expect to obtain *if the null hypothesis was true*, and the resulting SPD can be constructed using standard procedures. To account for variations arising from sampling error, this process is repeated many times. The resulting *distribution* of SPDs is then compared against the empirically observed one in two ways. The first consists of displaying the simulation envelope against the observed data and visually identifying regions of positive and negative deviations that represent time-interval where the density of radiocarbon dates was higher or lower than the one expected by the null model. The second consists of retrieving a single, global P-value based on a test statistic computed from the aggregate deviation from the simulation envelope (see Timpson et al. 2014 for details).

The MCMC approach effectively addresses two of the most problematic issues (i.e., sampling error and calibration effect) by emulating their consequences in the Monte-Carlo simulation routine. While there have been some minor modifications in the method (see for example the use of different algorithms for generating samples - see Crema and Bevan 2021), as well as some follow-up secondary analyses (e.g., Edinborough et al. 2017), the fundamental approach remains the same and is implemented in the R packages *rcarbon* (Crema and Bevan 2021) and *ADMUR* (Timpson et al. 2021).

The method described above is effectively a one-sample test where the observed SPD is compared against a user-defined theoretical model. In many situations, however, the key objective is to compare two or more SPDs to each other rather than against a theoretical model. Examples include the comparison of the population trajectory of two or more geographic regions (Shennan et al. 2013) or the relative proportion of different site types (e.g., monuments vs settlements; as in Collard et al. 2010) or dated samples (e.g. wild vs domesticated plants; as in Stevens and Fuller 2012). All these cases can be tackled using a randomisation test, which simply consists of: 1) assigning a *mark* to each radiocarbon date defining its membership to a particular set (e.g. region A and region B); 2) generating a separate SPD for each set; 3) randomly shuffling the *marks* assigned to the dates, and generating an SPD for each set again; 4) repeating the previous step multiple times; 5) comparing the observed SPD obtained in step 2 against the distribution of SPDs obtained in step 4 using a similar procedure to the one-sample Monte-Carlo method described above. Such *mark permutation test* (Crema et al. 2016; but see also Dye 1995 for a similar earlier application) provides a direct test on whether multiple SPDs have similar *shapes* and is

currently implemented in the *rcarbon* R package. Extensions of this approach include hot-spot analyses for detecting spatial heterogeneity in growth rates (Crema et al. 2017), and formal testing of resilience-resistance to external perturbation (Riris and de Souza 2021).

NHST approaches to the analysis of time-frequency data have successfully introduced a more robust inferential process that overcame many of the limitations imposed by simple visual assessments of SPDs. Whilst these advances are important steps forward; they also share the same kind of problems afflicting the NHST framework in general. Three of them are particularly noteworthy and deserve some careful consideration.

Firstly, the interpretation of P-values should account that these are both a function of sample and effect sizes. While I am not aware of any systematic survey on the misinterpretation of P-values in archaeology, review studies in other fields that employ statistical inference more routinely suggest that its definition and interpretation are often incorrect (e.g., Gliner et al. 2010, Greenland et al. 2016). A high P-value should not be interpreted as a goodness of fit of the radiocarbon record to the proposed null model, whilst low P-values can easily be obtained if there is a sufficiently large sample size, even if the effect size (i.e., the deviation from the null hypothesis) is comparatively small. The second point highlights the main inferential limitation of NHST, particularly when quantifiable estimates of effect sizes are not available, as in this case. Testing whether an observed SPD deviates from a particular exponential growth rate or determining whether two regions have different trajectories are examples of point hypotheses, i.e. a hypothesis that evaluates a *single* value. Strictly speaking, we *already know that the null hypothesis is incorrect* — an SPD would unlikely have exactly a particular exponential growth rate at its 7th decimal point, and two regions would never have perfectly identical population dynamics. What matters is how and how much the observed data deviates from a particular null hypothesis, and this is not something that can be inferred from P-values. Obtaining a statistically significant result might well just tell us only that we have a large number of radiocarbon dates in our databases.

Secondly, while the selection of the null hypothesis for permutation tests is typically straightforward, one-sample Monte-Carlo tests require a user-defined growth model. This means that depending on the choice of this null model, global P-values, as well as local positive and negative deviations from the simulation envelope, can vary. For example, using an exponential growth null model for radiocarbon frequency data characterised by a logistic growth would yield a negative deviation for time intervals where the population reached its carrying capacity. Similarly, large deviations from the null model during early sections of the window of analyses can lead to misleading signals in later portions even if the underlying shape of the SPDs are similar. Comparing rates of change of the SPDs can partly solve the problem (see e.g. Crema and Kobayashi 2020, Arroyo-Calin and Riris 2021), but clearly, positive and negative deviations should not be uncritically interpreted as signals of population boom and busts. It is also worth pointing out that some instances of local deviations are expected to be false positives (see Timpson et al 2021 for discussion), and as such, interpretation of these plots should only be made only if the global P-value suggests a rejection of the null hypothesis in the first place.

Thirdly, it should be noted that the one-sample Monte-Carlo method is designed to test the observed SPD against a particular parametrisation of a model. In other words, the question that is being asked is not whether a given data follows, for example, an exponential growth, but whether it follows an exponential growth with a *specific* growth rate r. It follows that rejecting a particular rate r does not necessarily imply that all exponential growth models are

rejected. In practice, however, one could test against the most probable value of *r* so that its rejection would imply the rejection of all other values of *r* and consequently the model as a whole. The selection of *r* (or any other parameters) is typically obtained by fitting a regression model to the observed SPD values. As discussed above (and explored in Carleton 2021), these estimates can be biased (see also Fig. 2). It is difficult to determine whether the impact of this discrepancy can have significant inferential consequences, and it is worth noting that the approach does not necessitate a workflow where the null model is based on the observed data. For example, Silva and Vander Linden (2017) examined SPDs of Neolithic Europe using the growth rate estimated from pre-existing Mesolithic populations, whilst Crema and Kobayashi (2020) have compared an SPD of the Jomon period in central Japan against a null model based on the fluctuations of independently dated pit-dwellings.

Model-fitting approaches

Both reconstructive and NHST approaches are commonly used as exploratory devices that provide the basis for developing more sophisticated explanatory models. These are, however, mostly limited to speculative statements that are rarely tested directly or formally compared against alternative hypotheses. The desire to move beyond this inferential framework has led to a steadily growing number of studies that have attempted to use SPDs in more ingenious ways. In many cases, however, this endeavour is being pursued by directly using SPDs *as the observed data,* effectively ignoring the potential bias of sampling error and calibration effects (see discussion above).

In 2021 alone, four different solutions have been developed to address these issues and provide a framework that can be used to fit putative growth models, infer their parameters, and carry out formal comparisons between competing hypotheses. While some of these approaches share similarities from a methodological standpoint, they are effectively distinct approaches with different accuracy, flexibility, and computational performance levels.

Carleton (2021) proposes a hierarchical Bayesian workflow named *Radiocarbon-dated Event Count model* (hereafter REC model), which models the radiocarbon record as a one-dimensional point process with a time-varying intensity parameter $\lambda(t)$. REC consists of fitting a hierarchical generalised linear model (GLM) that includes time as one of its covariates and optionally a set of additional independent variables (e.g., climate record). The key idea behind REC is to tackle the problem of chronological uncertainty by sampling *n* sets of random calendar dates from the calibrated distribution of each radiocarbon date and generating *n* vectors of count frequencies based on user-defined temporal bins. These sets of count data are then fitted using either a Poisson or Negative binomial regression. The hierarchical structure of REC ensures that the distribution of the n regression coefficients is directly modelled using Gaussian distributions, which moments are effectively the estimate and the associated uncertainty of our parameters of interest. Carleton tested the accuracy of the REC model by generating a simulated dataset with a known exponential growth rate and showed that although it fails to recover the correct value within its posterior range, it does offer a considerable improvement over the direct application of GLM on SPD values (figure 13, Carelton 2021, but see also Fig 2). The two main limitations of this approach are its high computational cost, which increases when the temporal resolution and the number of sampled sets of dates *n* are high, and the requirement for a comparatively large sample size. The latter point is intrinsically linked to the idea of using a count-based statistic where effectively the samples are not the observed number of dates but the number of temporal bins. It follows that an absence of dates in a particular bin could be evidence of low intensity or simply the effect of sampling error. In other words, the sampling procedures address the issue of chronological

uncertainty but not sampling error. When a larger number of radiocarbon dates is available, the potential bias in the output is reduced, but when sample sizes are small, one should interpret the estimates as descriptive statistics of the sample rather than inferred population parameters. Despite these shortcomings, the opportunity to directly integrate external covariates is appealing and has already led to its application in determining the role of climate change in the extinction of quaternary megafauna in North America (Stewart et al. 2021). A dedicated R package (*chronup*) with a revised method that addresses some of these concerns is currently being developed (see Carleton and Campbell 2021).

Porčić et al. (2021) have instead employed a generative inference approach where estimates are made by first simulating a large collection of SPDs with the same samples size as the observed data and using different "candidate" parameter combinations of a particular population model. These outputs are then individually compared to the observed SPD, and the parameter values used in the subset of simulations with the closes fit to this target are interpreted as an approximation of the estimate. This approach, known as approximate Bayesian computation (hereafter ABC), was initially developed in population genetics (Beaumont et al 2002) and has been successfully applied in different fields, including archaeology (Kovacevic et al. 2015, Crema et al. 2016, Carrignon et al. 2020). In the case of radiocarbon frequency data, the generative approach effectively solves the problem of sampling error and calibration effects following the same principles of the one-sample Monte-Carlo simulation method described above. The key difference is the definition of an initial prior distribution of possible parameter values from which these SPDs are simulated. Porčić et al. (2021) used a distance measure to evaluate the similarity between their candidate and observed SPDs, which they then used to define a subset of parameter combinations yielding the closest fit to data. These subsets are approximations of the posterior distribution for each of the model parameters. The most appealing feature of ABC is the great flexibility in defining the generative model, as evidenced by its recent application coupled with agent-based simulations (Carrignon et al. 2020). The already mentioned re-analyses of the radiocarbon record from Rapa Nui by Di Napoli et al. (2021) is an example that showcases how this approach can be used to fit complex ecological models such as logistic growths with time-varying and externally dependent carrying capacities. However, the flexibility of ABC is countered by the extreme computational cost required to obtain a sufficiently large number of posterior samples for an accurate and precise estimate of the parameter of interest. The development of more efficient algorithms (Beaumont 2019) is reducing this computational cost, but part of the issue is also dictated by the details of the simulation model itself. While there are no dedicated software packages for this approach either, both Porčić et al. 2021 and Di Napoli et al. 2021 provide R scripts that can be tailored to specific needs (see also the script used for Fig. 2 below).

ABC is typically employed in situations where the likelihood function of a particular model cannot be numerically computed and hence substituted by a large number of simulations and a measure of discrepancy between target and candidate. Numerical solutions of the likelihood function are available for common probability distributions, such as the uniform or the Gaussian, that are routinely employed in radiocarbon phase modelling (Buck et al. 1992). However, these probability distributions rarely represent suitable models of population change (but see the finite Gaussian mixture model discussed), particularly so when the latter is more complex, as in the example of the time-varying carrying capacity model described above. From a mathematical standpoint, the complexity arises because time is modelled as a continuum, and hence the likelihood is based on a probability density function. However, the likelihood calculation becomes trivial by treating time as discrete (i.e., using individual

766 calendar years as units) and using probability mass functions to model changes in the density
767 of radiocarbon dates over a given interval. Given a population growth model *m* with some
768 parameters θ_1, θ_2, … θ_k representing the probabilities of observing a radiocarbon date for each
769 *k* year within the window of analyses, the likelihood is equivalent to the product of the
770 probabilities of the observed events. For example, if our sample consists of three dates
771 x_1=3200, x_2=3300, and x_3=2800, and their probabilities for a particular growth model with
772 some defined parameter value *y* are π_1= 0.02, π_2=0.023 π_3=0.001, then the likelihood $L(\theta=y|$
773 $x_1,x_2,x_3)$ is equivalent to $\pi_1 \times \pi_2 \times \pi_3$, or 0.00000046. One can estimate the parameter *y*
774 yielding the highest likelihood given these three dates. The problem is that radiocarbon dates
775 are not single values but are instead described by a probability distribution that results from
776 its measurement error and the calibration process. Timpson et al. (2021). account for this
777 measurement error by basically calculating the scalar product between the model
778 probabilities and the probabilities from the calibrated dates. For example, suppose that x_1 now
779 has a probability of being equal to 3200 of 0.4 and a probability of being 3201 of 0.6. We
780 would update π_1 as (0.4 × probability of obtaining 3200 according to the model) ×
781 (0.6 × probability of getting 3201 according to the model).
782
783 This solution effectively enables the use of statistical tools based on likelihood estimation.
784 Model parameters can be inferred based on maximum likelihood, and alternative hypotheses
785 can be compared using information criteria. Because the calculation of the likelihood
786 function is effectively always the same, the model is also highly flexible. Any mathematical
787 model that can generate discrete probabilities within a bounded range of calendar years can
788 effectively be fitted with this approach. Timpson et al. (2021) make good use of this
789 flexibility and examined the radiocarbon record from the South American Arid Diagonal
790 using a continuous piecewise linear (CPL) model. The population growth model they employ
791 effectively consists of *n* linear segments and *n-1* hinge-points, which requires *2n-1*
792 parameters to be inferred. By using information criteria, they explore models with different
793 numbers of segments and show that 3-CPL (i.e., a three-segment model) provides the best fit
794 to the data, providing key information such as when major shifts in population growth rate
795 occurred in the South American Arid Diagonal region. This explicit model-based framework
796 also enables a more robust approach toward typical problems encountered in the analyses of
797 SPDs. For example, rather than applying a taphonomic "correction" to the observed summed
798 probabilities, ADMUR — the R package developed by Timpson et al. (2021) —allows for
799 the direct integration of the taphonomic loss model in the calculation of the likelihood and
800 consequently of the parameter estimates.
801
802 Crema and Shoda (2021) offer a Bayesian alternative to the solution developed by Timpson
803 et al. (2021). While the calculation of the likelihood function follows the same logic based
804 on the shift from probability density to probability mass functions, the modelling of
805 measurements errors and the possibility of using priors make their approach different. In
806 contrast to Timpson et al. (2021), their model considers calibrated probability distributions to
807 be posteriors that can be informed both from the individual observation (e.g., laboratory
808 measurement errors) and the higher-level model describing the variation in the density of
809 dates over time. This is conceptually the same approach used in Bayesian phase models
810 typically employed in software packages such as OxCal and BCal. As a result, the fitted
811 model estimates the population-level parameters (e.g., exponential growth rate) and the
812 posterior probability of each calibrated radiocarbon date. The second, and perhaps more
813 crucial difference, is the possibility to provide prior distributions to parameters of interest.
814 While strong priors and strict constraints as those occasionally implemented in Bayesian
815 phase models are unlikely to be useful in this context, the opportunity to use weakly

informative priors that can 'nudge' and reduce the possible range of parameters values (e.g. by, for example, reducing the probability of biologically implausible growth rates) can enormously help the inference process when sample sizes are limited, allowing researchers to implement stricter inclusion criteria for their available radiocarbon datasets.

The Bayesian nature of this inferential framework is particularly useful when the full extent of the uncertainty associated with the individual parameters is of interest. For example, in their case study, Crema and Shoda (2021) aimed to determine whether and when we observe a significant shift in population growth rate on the island of Kyushu in South-West Japan at the onset of the introduction of rice farming. They estimated this change-point to be around the 8th-7th century BC and used the earliest dated charred remains of rice to estimate a temporal lag of several centuries between the putative date of the introduction of farming and the timing of the demographic response. Similarly, Kim et al. (2021) investigated whether the population crash that occurred during the latter half of the Chulmun period (10,000 - 3,500 cal BP) resulted from mid-4th millennium climatic deterioration. To evaluate this hypothesis, they measured the temporal lag between the estimated start point of the population decline (as inferred from radiocarbon density) and the timing of abrupt changes estimated from Bayesian age-depth models of different proxies. Because both measures are characterised by chronological uncertainty, Kim et al. (2021) computed distributions of age differences from the estimated posteriors and calculated the probability that the population crash initiated *after* the climatic deterioration. While there were some differences, they showed that the probability of such an event was close to zero for at least two of the three proxies examined.

It is also worth noting that because the computational framework developed by Crema and Shoda (2021) is essentially just a Bayesian hierarchical model, there are opportunities to construct models that can benefit from more complex structures. For example, cross-regional analyses can employ a hierarchical structure where growth rates of each region are inferred via partial-pooling, i.e. informed to some extent by the growth rates of other regions. This provides more robust estimates compared to separated analyses for each region and, at the same time, offers opportunities to directly model inter-regional variability in growth rates.

The four model-fitting approaches described here all offer substantially more robust ways to infer model parameters compared to regression models directly applied to SPDs. Figure 2 shows the fitted value and the 95% confidence interval of the growth rate of two samples of 50 and 500 radiocarbon dates. The direct regression fit to the SPD fails to include the actual growth rate (dashed line), and the difference in sample size has minimal to no impact on the width of the confidence interval. Three out of the four approaches discussed here successfully manage to include the actual growth rate in their confidence intervals, with a wider confidence interval for the smaller data set. REC shows a mixed outcome instead, with the actual rate recovered only for the larger set and the smaller set yielding a narrower confidence interval than the other methods examined here. Similarly, although recovering the true parameter, the ABC approach performs less efficiently with substantially wider posterior intervals.

Model-fitting approaches also provide an important additional benefit of being able to formally compare alternative growth models against the observed data. For example, Timpson et al. (2021) employed Schwarz Criterion to determine the optimal number of hinges in their CPL model, and similarly, Di Napoli et al. (2021) used Bayes Factors to compare different ecological models, and Crema and Shoda (2021) used the Widely Applicable Information Criterion (WAIC) to determine whether a model with change point

provided more support in contrast to simple exponential growth. The epistemological shift from a single to multi-model inference is highly appealing, as it allows for formal grounds for the contrasting of competing hypotheses of demographic histories. There are, however, a couple of important issues to consider. Firstly, as mentioned earlier, the calculation of AIC and other information criteria on regression models directly applied to SPD values returns incorrect estimates. As such, those interested in this inferential framework will have to resort to one of the approaches described in this section. Secondly, multi-model inference provides only a relative measure of goodness-of-fit; the best model among the candidates can still be, in absolute terms, a terrible model. Timpson et al. (2021) tackle this problem by employing a goodness-of-fit test that is effectively equivalent to the one-sample Monte Carlo test discussed earlier, while both Crema and Shoda (2021) and Di Napoli et al. (2021) employ a graphical posterior predictive check. While the robustness of these sanity checks is limited with smaller sample sizes, they offer an important tool for the multi-model inference of radiocarbon frequency data.

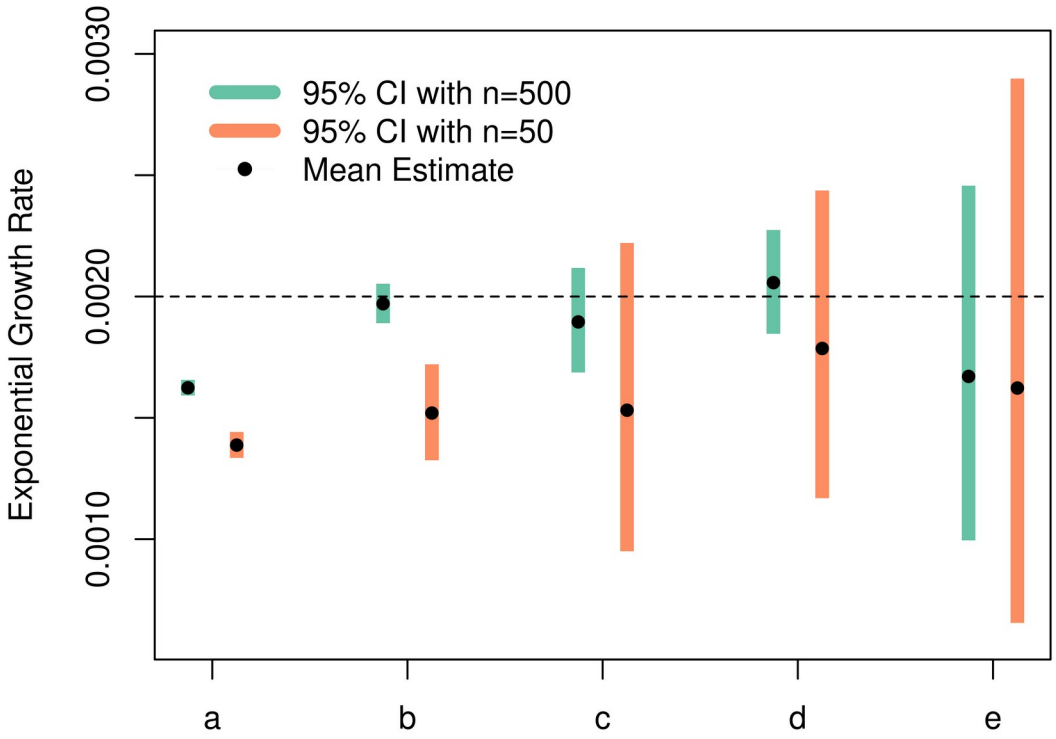

Figure 2 *Estimates and 95% confidence interval of a fitted exponential growth rate on a simulated dataset with two different sample sizes (n=50 and n=500) using: a) direct regression fit on the SPD; b) Bayesian radiocarbon-dated event count (REC) model; c) maximum likelihood fit via the ADMUR package; d) Bayesian hierarchical model via nimbleCarbon package; e) approximate Bayesian computation with rejection algorithm. Real growth rate is shown as a dashed line. R scripts and details required for generating the figures are available at https://github.com/ercrema/c14demoreview and archived on https://doi.org/10.5281/zenodo.6421345*

Where next?

The methodological review presented here showcases the growing range of analytical approaches designed to infer demographic changes from radiocarbon density data. While this trend is dictated by similar objectives and hence can be conceived as genuine alternatives, most of the methods discussed above were developed with different needs in mind. Some of the proposed solutions, particularly those grouped under *model-fitting* approaches, provide the foundation for developing bespoke analyses tailored to specific problems and questions arising from a given dataset. Others, such as those described here as *reconstructive* approaches, offer all-around solutions that are more suitable for an initial assessment of the available evidence. There is clearly no single go-to solution, and users should consider options according to their objectives. However, it is useful to highlight three recommendations that transcend these classifications and have often been raised by scholars who developed these techniques.

1. **SPD curves should never be exclusively interpreted from their visualisations nor *directly* used for statistical inference**. As mentioned repeatedly throughout this paper, the impact of sampling error and calibration effect is simply too significant to be ignored. Visual assessments of SPD can, however, provide important cues, particularly when dealing with broader-scale multi-millennial trends. As such, if the objective of the analysis is data description and exploration, the adoption of reconstructive approaches that visually provide an uncertainty envelope should be considered. While in some cases these methods might be too conservative and hide shorter scale fluctuations, they can avoid hasty conclusions based on little evidence.
2. **Consider running sensitivity analysis**. Many of the methods described above rely on some fine-tune settings where users are required to provide some numerical figures. These include, for example, binning window sizes for aggregating radiocarbon dates from the same site or bandwidth sizes in some Kernel Density Estimates. Although in some cases one can justify their choices, the relative impact of how changing these parameters affects the ultimate inference should be explored when possible (see for example Riris 2018, Feeser et al. 2019). Similarly, the inclusion or exclusion of a particular set of samples should be evaluated when possible. Such sensitivity analyses would reveal how changing these settings have no qualitative impact on the conclusion in the best-case scenario. Conversely, in the worst-case scenario, the ultimate results would depend on these decisions.
3. **Carry-out tactical models and *what-if* experiments.** Tactical models (Orton 1973, Lake 2014, Crema 2018) and *what-if experiments* (Buck and Meson 2015, Hinz 2020, Holland-Lulewicz and Ritchison 2021) are simulation techniques consisting of generating, in *silico,* artificial archaeological data under known conditions to determine the robustness of analytical techniques, explore the impact of particular biases, or estimate necessary sample sizes and guide data collection. These are powerful yet relatively underutilised tools that can enormously help in any statistical analysis. It is thus not surprising that these techniques have been used in radiocarbon density-based demographic research, either to establish the robustness of new or existing techniques (Contreas and Meadows 2014, Edinborough et al. 2017, Crema et al. 2017, Timpson et al. 2021, Carelton 2021, Price et al. 2021), question the impact of various forms of biases (e.g. Surovell and Brantingham 2007, Davies et al. 2016, Bevan and Crema 2021), or determine whether the available sample size is sufficient to recover putative demographic events (e.g. Hinz 2020, Crema and Shoda 2021). These techniques provide invaluable insights into the robustness of our analyses. They

can be tailored to the specific needs and challenges of particular contexts and even guide alternative solutions or more targeted future sampling strategies.

Some of these recommendations can be challenging to implement, particularly as they cannot be part of a generalised workflow and require a good understanding of the data set. Some techniques, such as ABC and OxCal's KDE, can also be computational too prohibitive to allow exhaustive sensitivity analyses or what-if experiments. Nonetheless, the benefit these tools provide is essential if we wish to make robust inferences about past population dynamics.

Despite these outstanding challenges, it is unquestionable that the appeal of radiocarbon-based population inference for comparative research remains. We are now able to, at least in principle, develop demographic models that are not limited to regional constraints of archaeological periodisations and start investigating common trajectories and detect anomalies. Several exciting studies have already started to move towards such a line of research, estimating benchmark figures of long-term population growth rates (Zahid et al. 2016) or identifying shared trajectories in their fluctuation at the global scale (Freeman et al. 2018). Similarly, continental-scale windows of analysis are revealing new insights and providing the grounds for developing new hypotheses (Shennan et al. 2013, Crema et al. 2017, Riris and Arroyo-Kalin 2019, Bird et al. 2020, Palmisano et al. 2021). While the methodological developments reviewed in this paper showcase the effort made by different research groups in addressing many of the concerns raised against early applications of radiocarbon density-based demographic inference, there is a clear trade-off between these large-scale comparative analyses and the inevitable increase in the number of biases that larger datasets entail. Local anomalies in the radiocarbon record might provide genuine insights that can help understand the demographic history of a particular region but might simply be the result of a spatially or chronologically structured bias. Incorrect inferences are inevitable, and the stakes can often be high. Still, the methodological advances made over the last few years and the high reward of expanding comparative demographic research in deep history suggest it is an endeavour well worth pursuing.

Acknowledgements

This work is the result of numerous discussions with many colleagues over the last few years to whom I am grateful. I would like to thank in particular Andrew Bevan, Adrian Timpson, Chris Carleton, and Mike Price for their insights on many of the topics discussed here and the two reviewers (Martin Hinz and an anonymous) for their feedback and comments on the manuscript. As is always the case, errors are my own.

Tables

Category	Name	Parameters	Parallel Processing	Computational Cost	Software	Reference
Reconstructive	Summed Probability Distribution	-	Yes	Low	*OxCal*, *rcarbon*, R scripts, *ADMUR*, *baydem*	Vv.Aa.
	Bayesian Gaussian Mixture	Number of Mixture Components (+); Number of MCMC iterations & burnin (+); Number of chains (+); Hyperpriors	Yes (MCMC chains)	High-Very High	*baydem*, *Bchron*	Price et al 2021
	Composite KDE	Kernel bandwidth size; Number of bootstrapped samples (+); Number resampled sets of calendar dates (+)	Yes	Low	*rcarbon*, R scripts	Brown 2017; McLaughlin 2018
	Bayesian KDE	-	No	High-Very High	*OxCal*	Bronk-Ramsey 2017
NHST	Monte-Carlo Summed Probability Distribution Method	Number of MC Simulations (+)	Yes	Medium	*rcarbon*, *ADMUR*	Shennan et al 2013; Timpson et al 2014; etc.
	Mark Permutation Test	Number of Permutations (+)	Yes	Low	*rcarbon*	Crema et al 2016; etc.
	Spatial Permutation Test	Number of Permutations (+)	Yes	Low	*rcarbon*	Crema et al 2017; etc.
	Point to Point Post-Hoc Test	Number of MC Simulations (+)	Yes	Low	*rcarbon*, R scripts	Edinborough et al 2017; see also Riris and De Souza 2021
Model Fitting	Approximate Bayesian Computation	Model; Priors; Number of Simulations (+); ABC algorithm; Tolerance level (-)	Yes	Very High	R scripts	Porčić et al 2021; Di Napoli et al 2021
	Bayesian Radiocarbon Event-Count Model	Model; Priors; Resolution of the temporal bins (-); Number resampled sets of calendar dates (+); Number of MCMC iterations & burn-in (+); Number of Chains (+)	Yes (MCMC chains)	High	R scripts; *chronup*	Carelton 2021; Stewart et al 2021
	Bayesian Hierarchical Model with Measurement Error	Model; Priors; Number of MCMC iterations & burnin (+); Number of chains (+)	Yes (MCMC chains)	High	*nimbleCarbon*	Crema and Shoda 2021; Kim et al 2021; Riris and De Souza 2021
	Maximum Likelihood Fitting	Model; Number of MCMC iterations (+)	No	High	*ADMUR*	Timpson et al 2021

Table 1. *Summary of statistical techniques for inferring past demography from radiocarbon frequency data.* Plus/minus signs in the *Parameters fields indicated whether preferences are for larger (+) or smaller (-) settings. Computational costs are indicative as it depends on sample size and availability of parallel processing: Low (<< 1 hour); Medium (~ several hours); High (~ 24 hours); Very High (~ several days*).

Compliance with Ethical Standards:

Funding: This work was funded by a Philip Leverhulme Prize (#PLP-2019-304).
Conflict of Interest: The author declares that he has no conflict of interest.

References

Ahn, S.-M., Hwang, J.H., (2015). Temporal fluctuation of human occupation during the 7th–3rd millennium cal BP in the central-western Korean Peninsula. *Quaternary International*, Quaternary Studies in Korea III: Contents and characteristics of paleoclimatology/paleoceanography studies in and around Korea, 384, 28–36. https://doi.org/10.1016/j.quaint.2015.04.038

Ames, K.M., (1991). The archaeology of the Longue Durée: Temporal and Spatial Scale in the Evolution of Social Complexity in Southern Northwest Coast. *Antiquity,* 65, 935–945.

Anscombe, F.J., (1973). Graphs in Statistical Analysis. *The American Statistician,* 27, 17–21. https://doi.org/10.1080/00031305.1973.10478966

Arroyo-Kalin, M., Riris, P., (2021). Did pre-Columbian populations of the Amazonian biome reach carrying capacity during the Late Holocene? *Philosophical Transactions of the Royal Society B: Biological Sciences,* 376, 20190715. https://doi.org/10.1098/rstb.2019.0715

Attenbrow, V., Hiscock, P., (2015). Dates and demography: are radiometric dates a robust proxy for long-term prehistoric demographic change? *Archaeology in Oceania,* 50, 30–36. https://doi.org/10.1002/arco.5052

Baxter, M.J., Cool, H.E.M., (2016). Reinventing the wheel? Modelling temporal uncertainty with applications to brooch distributions in Roman Britain. *Journal of Archaeological Science* 66, 120–127. https://doi.org/10.1016/j.jas.2015.12.007

Beaumont, M.A., Zhang, W., Balding, D.J., (2002). Approximate Bayesian Computation in Population Genetics. *Genetics* 162, 2025–2035.

Beaumont, M.A., (2019). Approximate Bayesian Computation. *Annual Review of Statistics and Its Application* 6, 379–403. https://doi.org/10.1146/annurev-statistics-030718-105212

Becerra-Valdivia, L., Leal-Cervantes, R., Wood, R., Higham, T., (2020). Challenges in sample processing within radiocarbon dating and their impact in 14C-dates-as-data studies. *Journal of Archaeological Science*, 113, 105043. https://doi.org/10.1016/j.jas.2019.105043

Bevan, A., Colledge, S., Fuller, D., Fyfe, R., Shennan, S., Stevens, C., (2017). Holocene fluctuations in human population demonstrate repeated links to food production and climate. *PNAS* ,114, E10524–E10531. https://doi.org/10.1073/pnas.1709190114

Bevan, A., Crema, E.R., (2021). Modifiable reporting unit problems and time series of long-term human activity. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 376, 20190726. https://doi.org/10.1098/rstb.2019.0726

Bird, D., Freeman, J., Robinson, E., Maughan, G., Finley, J.B., Lambert, P.M., Kelly, R.L., (2020). A first empirical analysis of population stability in North America using radiocarbon records. *The Holocene*, 30, 1345–1359. https://doi.org/10.1177/0959683620919975

Bird, D., Miranda, L., Vander Linden, M., Robinson, E., Bocinsky, R.K., Nicholson, C., Capriles, J.M., Finley, J.B., Gayo, E.M., Gil, A., d'Alpoim Guedes, J., Hoggarth, J.A., Kay, A., Loftus, E., Lombardo, U., Mackie, M., Palmisano, A., Solheim, S., Kelly, R.L., Freeman, J., (2022). p3k14c, a synthetic global database of archaeological radiocarbon dates. *Scientific Data*, 9, 27. https://doi.org/10.1038/s41597-022-01118-7

Blackwell, P.G., Buck, C.E., (2003). The Late Glacial human reoccupation of north-western Europe: new approaches to space-time modelling. *Antiquity*, 77, 232–240.

Blockley, S.P.E., Donahue, R.E., Pollard, A.M., (2000). Radiocarbon calibration and Late Glacial occupation in northwest Europe. *Antiquity*, 74, 112–119. https://doi.org/10.1017/S0003598X00066199

Bluhm, L.E., Surovell, T.A., (2019). Validation of a global model of taphonomic bias using geologic radiocarbon ages. *Quaternary Research*, 91, 325–328. https://doi.org/10.1017/qua.2018.78

Bronk Ramsey, C., (2017). Methods for Summarizing Radiocarbon Datasets. *Radiocarbon,* 59, 1809–1833. https://doi.org/10.1017/RDC.2017.108

Broughton, J.M., Weitzel, E.M., (2018). Population reconstructions for humans and megafauna suggest mixed causes for North American Pleistocene extinctions. *Nature Communications*, 9, 5441. https://doi.org/10.1038/s41467-018-07897-1

Brown, W.A., (2015). Through a filter, darkly: population size estimation, systematic error, and random error in radiocarbon-supported demographic temporal frequency analysis. *Journal of Archaeological Science*, 53, 133–147. https://doi.org/10.1016/j.jas.2014.10.013

Brown, W.A., (2017). The past and future of growth rate estimation in demographic temporal frequency analysis: Biodemographic interpretability and the ascendance of dynamic growth models. *Journal of Archaeological Science*, 80, 96–108. https://doi.org/10.1016/j.jas.2017.02.003

Buck, C.E., Litton, C.D., Smith, A.F.M., (1992). Calibration of radiocarbon results pertaining to related archaeological events. *Journal of Archaeological Science*, 19, 497–512. https://doi.org/10.1016/0305-4403(92)90025-X

Buck, C.E., Meson, B., (2015). On being a good Bayesian. *World Archaeology*, 47, 567–584. https://doi.org/10.1080/00438243.2015.1053977

Carleton, W.C., (2021). Evaluating Bayesian Radiocarbon-dated Event Count (REC) models for the study of long-term human and environmental processes. *Journal of Quaternary Science*, 36, 110–123. https://doi.org/10.1002/jqs.3256

Carleton, W.C., Campbell D.A. (2021). Improved parameter estimation and uncertainty propagation in Bayesian Radiocarbon-Dated Event Count (REC) models. OSF Preprint https:// https://osf.io/56dbt/

Carleton, W.C., Groucutt, H.S., (2021). Sum things are not what they seem: Problems with point-wise interpretations and quantitative analyses of proxies based on aggregated radiocarbon dates. *The Holocene*, 31, 630–643. https://doi.org/10.1177/0959683620981700

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. https://doi.org/10.1371/journal.pone.0240414

Chatters, J.C., (1995). Population Growth, Climatic Cooling, and the Development of Collector Strategies on the Southern Plateau, Western North America. *Journal of World Prehistory*, 9, 341–400.

Chaput, M.A., Kriesche, B., Betts, M., Martindale, A., Kulik, R., Schmidt, V., Gajewski, K., (2015). Spatiotemporal distribution of Holocene populations in North America. *PNAS*, 112, 12127–12132. https://doi.org/10.1073/pnas.1505657112

Chaput, M.A., Gajewski, K., (2016). Radiocarbon dates as estimates of ancient human population size. *Anthropocene*, 15, 3–12. https://doi.org/10.1016/j.ancene.2015.10.002

Collard, M., Edinborough, K., Shennan, S., Thomas, M.G., (2010). Radiocarbon evidence indicates that migrants introduced farming to Britain. *Journal of Archaeological Science*, 37, 866–870. https://doi.org/10.1016/j.jas.2009.11.016

Collins-Elliott, S.A., (2019). Quantifying artefacts over time: Interval estimation of a Poisson distribution using the Jeffreys prior. *Archaeometry*, 61, 1207–1222. https://doi.org/10.1111/arcm.12481

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

Crema, E.R., (2012). Modelling Temporal Uncertainty in Archaeological Analysis. *Journal of Archaeological Method and Theory*, 19, 440–461.

Crema, E.R., (2018). Statistical inference and archaeological simulations. *The SAA Archaeological Record*, 18, 20–23.

Crema, E.R., (2020). Non-Stationarity and Local Spatial Analysis, in: Gillings, M., Hacıgüzeller, P., Lock, G. (Eds.), *Archaeological Spatial Analysis*. Routledge, London, pp. 155–168.

Crema, E.R., Habu, J., Kobayashi, K., Madella, M., (2016). Summed Probability Distribution of 14 C Dates Suggests Regional Divergences in the Population Dynamics of the Jomon Period in Eastern Japan. *PLOS ONE*, 11, e0154809. https://doi.org/10.1371/journal.pone.0154809

Crema, E.R., Kandler, A., Shennan, S., (2016). Revealing patterns of cultural transmission from frequency data: equilibrium and non-equilibrium assumptions. *Scientific Reports*, 6, 39122. https://doi.org/10.1038/srep39122

Crema, E.R., Bevan, A., Shennan, S., (2017). Spatio-temporal approaches to archaeological radiocarbon dates. *Journal of Archaeological Science*, 87, 1–9. https://doi.org/10.1016/j.jas.2017.09.007

Crema, E.R., Kobayashi, K., (2020). A multi-proxy inference of Jōmon population dynamics using Bayesian phase models, residential data, and summed probability distribution of 14C dates. *Journal of Archaeological Science,* 117, 105136. https://doi.org/10.1016/j.jas.2020.105136

Crema, E.R., Bevan, A., (2021). Inference from large sets of radiocarbon dates: software and methods. *Radiocarbon,* 63, 23–39. https://doi.org/10.1017/RDC.2020.95

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. https://doi.org/10.1371/journal.pone.0251695

Davies, B., Holdaway, S.J., Fanning, P.C., (2016). Modelling the palimpsest: An exploratory agent-based model of surface archaeological deposit formation in a fluvial arid Australian landscape. *The Holocene*, 26, 450–463.

Di Napoli, 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). *Nature Communications*, 12, 3939. https://doi.org/10.1038/s41467-021-24252-z

Downey, S.S., Haas, W.R., Shennan, S.J., (2016). European Neolithic societies showed early warning signals of population collapse. *PNAS*, 113, 9751–9756. https://doi.org/10.1073/pnas.1602504113

Drennan, Robert D., Berry, A.C., Peterson, C.E., (2015). *Regional Settlement Demography in Archaeology*, Principles of Archaeology. Eliot Werner Publications, New York.

Dye, T., (1995). Comparing 14C Histograms: An Approach Based on Approximate Randomization Techniques. *Radiocarbon*. 37, 851–859. https://doi.org/10.1017/S0033822200014934

Dye, T.S., (2016). Long-term rhythms in the development of Hawaiian social stratification. *Journal of Archaeological Science*, 71, 1–9. https://doi.org/10.1016/j.jas.2016.05.006

Dye, T., Komori, E., (1992). A Pre-Censal Population History of Hawai'i. *New Zealand Journal of Archaeology*, 14, 113–128.

Edinborough, K., Porčić, M., Martindale, A., Brown, T.J., Supernant, K., Ames, K.M., (2017). Radiocarbon test for demographic events in written and oral history. *PNAS*, 114, 12436–12441. https://doi.org/10.1073/pnas.1713012114

Erlandson, J., Crowell, A., Wooley, C., Haggarty, J., (1992). Spatial and Temporal Patterns in Alutiiq Paleodemography. *Arctic Anthropology*, 29, 42–62.

Fernandes, R., Millard, A.R., Brabec, M., Nadeau, M.-J., Grootes, P., (2014). Food Reconstruction Using Isotopic Transferred Signals (FRUITS): A Bayesian Model for Diet Reconstruction. *PLOS ONE* 9, e87436. https://doi.org/10.1371/journal.pone.0087436

Fernández-López de Pablo, J., Gutiérrez-Roig, M., Gómez-Puche, M., McLaughlin, R., Silva, F., Lozano, S., (2019). Palaeodemographic modelling supports a population bottleneck during the Pleistocene-Holocene transition in Iberia. *Nature Communications,* 10, 1872. https://doi.org/10.1038/s41467-019-09833-3

Feeser, I., Dörfler, W., Kneisel, J., Hinz, M., Dreibrodt, S., (2019). Human impact and population dynamics in the Neolithic and Bronze Age: Multi-proxy evidence from north-western Central Europe. *The Holocene*, 29, 1596–1606. https://doi.org/10.1177/0959683619857223

Freeman, J., Hard, R.J., Mauldin, R.P., Anderies, J.M., (2021). Radiocarbon data may support a Malthus-Boserup model of hunter-gatherer population expansion. *Journal of Anthropological Archaeology*, 63, 101321. https://doi.org/10.1016/j.jaa.2021.101321

Geyh, M.A., (1980). Holocene Sea-Level History: Case Study of the Statistical Evaluation of 14C Dates. *Radiocarbon,* 22, 695–704. https://doi.org/10.1017/S0033822200010067

Gleeson, P., McLaughlin, R., (2021). Ways of death: cremation and belief in first-millennium AD Ireland. *Antiquity*, 95, 382–399. https://doi.org/10.15184/aqy.2020.251

Gliner, J.A., Leech, N.L., Morgan, G.A., (2002). Problems With Null Hypothesis Significance Testing (NHST): What Do the Textbooks Say? *The Journal of Experimental Education*, 71, 83–92. https://doi.org/10.1080/00220970209602058

Greenland, S., Senn, S.J., Rothman, K.J., Carlin, J.B., Poole, C., Goodman, S.N., Altman, D.G., (2016). Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. *European Journal of Epidemiology*, 31, 337–350. https://doi.org/10.1007/s10654-016-0149-3

Guilderson, T.P., Reimer, P.J., Brown, T.A., (2005). The Boon and Bane of Radiocarbon Dating. *Science,* 307, 362–364. https://doi.org/10.1126/science.1104164

Haslett, J., Parnell, A., (2008). A simple monotone process with application to radiocarbon-dated depth chronologies. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 57, 399–418. https://doi.org/10.1111/j.1467-9876.2008.00623.x

Heidenreich, N.-B., Schindler, A., Sperlich, S., (2013). Bandwidth selection for kernel density estimation: a review of fully automatic selectors. *AStA Advances in Statistical Analysis*, 97, 403–433. https://doi.org/10.1007/s10182-013-0216-y

Hinz, M., (2020). Sensitivity of radiocarbon sum calibration. *Journal of computer applications in archaeology*, 3, 238–252.

Holland-Lulewicz, J., Ritchison, B.T., (2021). How Many Dates Do I Need?: Using Simulations to Determine Robust Age Estimations of Archaeological Contexts. *Advances in Archaeological Practice*, 9, 272–287. https://doi.org/10.1017/aap.2021.10

Housley, R.A., Gamble, C.S., Street, M., Pettitt, P., (1997). Radiocarbon evidence for the Lateglacial Human Recolonisation of Northern Europe. *Proceedings of the Prehistoric Society*, 63, 25–54. https://doi.org/10.1017/S0079497X0000236X

Johnson, I., (2004). Aoristic Analysis: seeds of a new approach to mapping archaeological distributions through time., in: Ausserer, K.F., ¨rner, W.B., Goriany, M., ckl, L.K.-V.¨(Eds.), *[Enter the Past] the E-Way into the Four Dimensions of Cultural Heritage: CAA2003*. BAR International Series 1227. Archaeopress, Oxford, pp. 448–452.

Kelly, R.L., Surovell, T.A., Shuman, B.N., Smith, G.M., (2013). A continuous climatic impact on Holocene human population in the Rocky Mountains. *PNAS,* 110, 443–447. https://doi.org/10.1073/pnas.1201341110

Kim, H., Lee, G.-A., Crema, E.R., (2021). Bayesian analyses question the role of climate in Chulmun demography. *Scientific Reports*, 11, 23797. https://doi.org/10.1038/s41598-021-03180-4

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.

Lake, M.W., (2014). Trends in Archaeological Simulation. *Journal of Archaeological Method and Theory*, 21, 258–287. https://doi.org/10.1007/s10816-013-9188-1

Lima, M., Gayo, E.M., Latorre, C., Santoro, C.M., Estay, S.A., Cañellas-Boltà, N., Margalef, O., Giralt, S., Sáez, A., Pla-Rabes, S., Chr. Stenseth, N., (2020). Ecology of the collapse of Rapa Nui society. *Proceedings of the Royal Society B: Biological Sciences*, 287, 20200662. https://doi.org/10.1098/rspb.2020.0662

Lucarini, G., Wilkinson, T., Crema, E.R., Palombini, A., Bevan, A., Broodbank, C., (2020). The MedAfriCarbon Radiocarbon Database and Web Application. Archaeological Dynamics in Mediterranean Africa, ca. 9600–700 BC. *Journal of Open Archaeology Data,* 8, 1. https://doi.org/10.5334/joad.60

Martínez-Grau, H., Morell-Rovira, B., Antolín, F., (2021). Radiocarbon Dates Associated to Neolithic Contexts (Ca. 5900 – 2000 Cal BC) from the Northwestern Mediterranean Arch to

the High Rhine Area. *Journal of Open Archaeology Data,* 9, 1. https://doi.org/10.5334/joad.72

Manning, K., Colledge, S., Crema, E., Shennan, S., Timpson, A., (2016). The Cultural Evolution of Neolithic Europe. EUROEVOL Dataset 1: Sites, Phases and Radiocarbon Data. *Journal of Open Archaeology Data,* 5. https://doi.org/10.5334/joad.40

Manning, K., Timpson, A., (2014). The demographic response to Holocene climate change in the Sahara. *Quaternary Science Reviews*, 101, 28–35. https://doi.org/10.1016/j.quascirev.2014.07.003

McLaughlin, T.R., (2019). On Applications of Space–Time Modelling with Open-Source 14C Age Calibration. *Journal of Archaeological Method and Theory*, 26, 479–501. https://doi.org/10.1007/s10816-018-9381-3

Michczyńska, D.J., Pazdur, A., (2004). Shape Analysis of Cumulative Probability Density Function of Radiocarbon Dates Set in the Study of Climate Change in the Late Glacial and Holocene. *Radiocarbon,* 46, 733–744. https://doi.org/10.1017/S0033822200035773

Oh, Y., Conte, M., Kang, S., Kim, J., Hwang, J., (2017). Population Fluctuation and the Adoption of Food Production in Prehistoric Korea: Using Radiocarbon Dates as a Proxy for Population Change. *Radiocarbon,* 59, 1761–1770. https://doi.org/10.1017/RDC.2017.122

Orton, C., (1973). The tactical use of models in archaeology - the SHERD project, in: Renfrew, C. (Ed.), *The Explanation of Culture Change*. Duckworth, London, pp. 137–139.

Palmisano, A., Bevan, A., Shennan, S., (2017). Comparing archaeological proxies for long-term population patterns: An example from central Italy. *Journal of Archaeological Science,* 87, 59–72. https://doi.org/10.1016/j.jas.2017.10.001

Palmisano, A., Lawrence, D., de Gruchy, M.W., Bevan, A., Shennan, S., (2021). Holocene regional population dynamics and climatic trends in the Near East: A first comparison using archaeo-demographic proxies. *Quaternary Science Reviews,* 252, 106739. https://doi.org/10.1016/j.quascirev.2020.106739

Porčić, M., Blagojević, T., Pendić, J., Stefanović, S., (2021). The Neolithic Demographic Transition in the Central Balkans: population dynamics reconstruction based on new radiocarbon evidence. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 376, 20190712. https://doi.org/10.1098/rstb.2019.0712

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. *Journal of Archaeological Science*, 135, 105473. https://doi.org/10.1016/j.jas.2021.105473

Rick, John W., (1987). Dates as Data: An Examination of the Peruvian Radiocarbon Record. *American Antiquity,* 52, 55–73.

Riris, P., (2018). Dates as data revisited: A statistical examination of the Peruvian preceramic radiocarbon record. *Journal of Archaeological Science,* 97, 67–76. https://doi.org/10.1016/j.jas.2018.06.008

Riris, P., Arroyo-Kalin, M., (2019). Widespread population decline in South America correlates with mid-Holocene climate change. *Scientific Reports,* 9, 6850. https://doi.org/10.1038/s41598-019-43086-w

Riris, P., de Souza, J.G., (2021). Formal Tests for Resistance-Resilience in Archaeological Time Series. *Frontiers in Ecology and Evolution,* 9, 906. https://doi.org/10.3389/fevo.2021.740629

Seidensticker, D., Hubau, W., Verschuren, D., Fortes-Lima, C., de Maret, P., Schlebusch, C.M., Bostoen, K., (2021). Population collapse in Congo rainforest from 400 CE urges reassessment of the Bantu Expansion. *Science Advances,* 7, eabd8352. https://doi.org/10.1126/sciadv.abd8352

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

Silva, F., Vander Linden, M., (2017). Amplitude of travelling front as inferred from 14 C predicts levels of genetic admixture among European early farmers. *Scientific Reports,* 7, 11985. https://doi.org/10.1038/s41598-017-12318-2

Stevens, C.J., Fuller, D.Q., (2012). Did Neolithic farming fail? The case for a Bronze Age agricultural revolution in the British Isles. *Antiquity,* 86, 707–722.

Stewart, M., Carleton, W.C., Groucutt, H.S., (2021). Climate change, not human population growth, correlates with Late Quaternary megafauna declines in North America. *Nature Communications,* 12, 965. https://doi.org/10.1038/s41467-021-21201-8

Surovell, T.A., Brantingham, P.J., (2007). A note on the use of temporal frequency distributions in studies of prehistoric demography. *Journal of Archaeological Science,* 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. *Journal of Archaeological Science,* 36, 1715–1724.

Tallavaara, M., Jørgensen, E.K., (2021). Why are population growth rate estimates of past and present hunter–gatherers so different? *Philosophical Transactions of the Royal Society B: Biological Sciences,* 376, 20190708. https://doi.org/10.1098/rstb.2019.0708

Timpson, A., Colledge, S., Crema, E., Edinborough, K., Kerig, T., Manning, K., Thomas, M.G., 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. https://doi.org/10.1016/j.jas.2014.08.011

Timpson, A., Barberena, R., Thomas, M.G., Méndez, C., Manning, K., (2021). Directly modelling population dynamics in the South American Arid Diagonal using 14C dates.

Philosophical Transactions of the Royal Society B: Biological Sciences, 376, 20190723. https://doi.org/10.1098/rstb.2019.0723

Torfing, T., (2015). Neolithic population and summed probability distribution of 14C-dates. *Journal of Archaeological Science,* 63, 193–198. https://doi.org/10.1016/j.jas.2015.06.004

Tremayne, A.H., Winterhalder, B., (2017). Large mammal biomass predicts the changing distribution of hunter-gatherer settlements in mid-late Holocene Alaska. *Journal of Anthropological Archaeology,* 45, 81–97. https://doi.org/10.1016/j.jaa.2016.11.006

Vander Linden, M., (2019). Le rôle des diagnostics dans les recherches à visée synthétique : exemples pré- et protohistoriques, in: Flotté, D.D., Marcigny, C. (Eds.), *Le diagnostic comme outil de recherche : actes du 2e séminaire scientifique et technique de l'Inrap*. https://doi.org/10.34692/rrgd-xn86

Wang, C., Lu, H., Zhang, J., Gu, Z., He, K., (2014). Prehistoric demographic fluctuations in China inferred from radiocarbon data and their linkage with climate change over the past 50,000 years. *Quaternary Science Reviews,* 98, 45–59. https://doi.org/10.1016/j.quascirev.2014.05.015

Ward, I., Larcombe, P., (2021). Sedimentary unknowns constrain the current use of frequency analysis of radiocarbon data sets in forming regional models of demographic change. *Geoarchaeology,* 36, 546–570. https://doi.org/10.1002/gea.21837

Ward, G.K., Wilson, S.R., (1978). Procedures for Comparing and Combining Radiocarbon Age Determinations: A Critique. *Archaeometry,* 20, 19–31. https://doi.org/10.1111/j.1475-4754.1978.tb00208.x

Weninger, B., Clare, L., Jöris, O., Jung, R., Edinborough, K., (2015). Quantum theory of radiocarbon calibration. *World Archaeology,* 47, 543–566. https://doi.org/10.1080/00438243.2015.1064022

White, A.J., Stevens, L.R., Lorenzi, V., Munoz, S.E., Lipo, C.P., Schroeder, S., (2018). An evaluation of faecal stanols as indicators of population change at Cahokia, Illinois. *Journal of Archaeological Science*, 93, 129–134. https://doi.org/10.1016/j.jas.2018.03.009

Williams, A.N., (2012). The use of summed radiocarbon probability distributions in archaeology: a review of methods. *Journal of Archaeological Science,* 39, 578–589.

Zahid, H.J., Robinson, E., Kelly, R.L., (2016). Agriculture, population growth, and statistical analysis of the radiocarbon record. *PNAS* 113, 931–935. https://doi.org/10.1073/pnas.1517650112