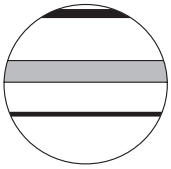
Check for updates



Research Paper

The Holocene
1–14
© The Author(s) 2020

Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0959683620981700
journals.sagepub.com/home/hol

Sum things are not what they seem: Problems with point-wise interpretations and quantitative analyses of proxies based on aggregated radiocarbon dates

W. Christopher Carleton[1] and **Huw S Groucutt**[1,2,3]

Abstract

Radiocarbon-date assemblages are commonly used as proxies for past human and environmental phenomena. Prominent examples of target phenomena include past population levels and sea level fluctuations. These processes are thought to have affected the amount of organic carbon deposited into the archaeological and/or palaeoenvironmental record. Time-series representing through-time fluctuations in the frequency of radiocarbon samples are, therefore, often used as proxies for such processes. However, there are critical problems with using radiocarbon "dates-as-data" in point-wise comparisons and these problems have gone largely underappreciated. The key problem is that the established proxies are easily misinterpreted. They conflate process variation and chronological uncertainty, which makes them unsuitable for point-wise comparisons aimed at identifying rates of change, comparing variables directly, or estimating parameters in regression models. Here we explore the interpretive and analytical problems in detail in an effort to raise awareness and promote skepticism about the use of the established proxies in point-wise comparisons. We also provide suggestions for future research and point to potential methodological alternatives that may improve the viability of dates-as-data approaches.

Keywords

archaeological proxies, palaeoclimatology, palaeodemography, palaeoenvironmental proxies, radiocarbon dates, summed probability density functions

Received 24 September 2020; revised manuscript accepted 16 November 2020

Introduction

Proxies based on aggregated radiocarbon-dates are popular in archaeological and palaeoenvironmental research. They are created by estimating through-time changes in the amount of radiocarbon deposited into the archaeological and palaeoenvironmental records. These time-series of radiocarbon amounts are thought to reflect target processes including, for example, population levels in archaeological research (e.g. d'Alpoim Guedes et al., 2016; Edinborough et al., 2017; Riede, 2009; Shennan, 2009) or past climate changes in palaeoenvironmental research (e.g. Bleicher, 2013; Thorndycraft and Benito, 2006; Turney and Brown, 2007). Widespread adoption of the proxies is having a significant impact on our understanding of both human–environment dynamics and climate processes.

The idea that radiocarbon dates could be used for something other than chronological control appears to have first been published in a 1969 paper by palaeoclimatologist Mebus Geyh. The paper reports a study of Holocene sea level changes along the North Sea coast with the help of "statistischen Auswertung von ^{14}C-Daten" ("statistical evaluation of ^{14}C data") (Geyh, 1969). In the paper, Geyh argued that the amount of radiocarbon (effectively, number of dated samples) in different layers of sediment from around the North Sea could be used as a proxy for past through-time fluctuations in sea level. His argument was based on the relationship between marine ingression, wetland formation, peat deposits, and carbon preservation. When sea levels rise, the argument goes, coastal areas experience an increase in wetland formation. This increase in turn leads to the formation of peat and, as a consequence, increased carbon preservation in sediments. Thus, Geyh observed, sediment cores will contain more carbon in layers that formed during periods of high sea level. Using this logic, he proposed a method for turning a set of radiocarbon dates into a proxy for past sea-level changes. The approach entailed approximating the Gaussian radiocarbon-date distributions with step-functions and then summing those functions (see Figure 1). Using this approach, Geyh created several sea-level proxies based on radiocarbon samples from coastal sediment cores around the North Sea. Each proxy corresponded to a region within the wider study area. He then compared the regional proxies and identified peaks in the proxies that appeared to be contemporaneous. These coeval peaks, he claimed, demonstrated sea-level change in the North Sea was regionally synchronous. Geyh then went on to use the technique in several more studies of Holocene climate change (e.g. Geyh, 1971, 1980).

[1]Extreme Events Research Group, Max Planck Institutes for Chemical Ecology, The Science of Human History, and Biogeochemistry, Germany
[2]Department of Archaeology, Max Planck Institute for The Science of Human History, Germany
[3]Institute of Prehistoric Archaeology, University of Cologne, Germany

Corresponding author:
W. Christopher Carleton, Extreme Events Research Group, Max Planck Institutes for Chemical Ecology, The Science of Human History, and Biogeochemistry, Hans Knöll Street 8, Jena 07745, Germany.
Email: wcarleton@ice.mpg.de

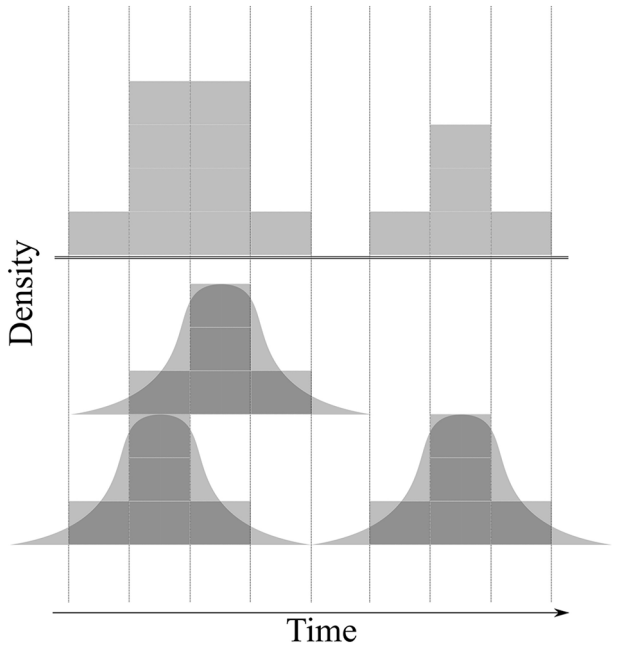


Figure 1. Step function summing technique developed by Geyh (1969) . Each radiocarbon-date density (represented by smooth Gaussian distributions in this image) is first approximated by a step function (represented here by blocks). Then, the height of each step function is summed for each interval of time to produce the summed radiocarbon-date curve, which is represented in this image by the blocky histogram-like figure above the double-line.

Independently, it seems, a similar notion occurred to archaeologists in the early 1970s. Rather than summing step-functions, though, early attempts to count radiocarbon-dated events involved binning median dates into temporal frequency histograms. Janette Deacon (1974), for example, used such a histogram as a proxy for fluctuations in human population levels in South Africa over the last 20,000 years. Deacon collected a database of 200 radiocarbon dates from South African archaeological sites and binned them into 1000-year time bins. The resulting distribution appeared to be dramatically increasing toward the present, which she argued indicated that population levels had increased over that period. Her reasoning was based on the idea that more radiocarbon dated sites in a given time/place indicated more people were present in that time/place. Other studies employed a similar approach with variations in sample size and temporal resolution (bin width) (e.g. Wendland and Bryson, 1974).

A few years later, in 1977, Garry Law suggested a methodological improvement (Black and Green, 1977). In a short contribution to a paper by Black and Green (1977) reporting radiocarbon sample data from the Solomon Islands, Law argued that binning median dates into a histogram failed to account for the probabilistic nature of radiocarbon assays. He proposed instead that the dates should be treated like probability densities and then added together. The approach was coincidentally similar to Geyh's, but involved a much higher temporal resolution and, consequently, produced smoother curves.

In the following decade, proxies based on radiocarbon dates appeared in several studies that would catalyze growth in the method's popularity among archaeologists. Michael Berry (1982) examined the temporal distribution of radiocarbon dates from the southwestern US to infer Ancestral Puebloan (Anasazi) population history from roughly 2200 to 1300 years BP. He argued that changes through time in numbers of dates from known sites tracked changes in population size and territorial abandonment. Declines in numbers of dates, he argued, corresponded to important climatic shifts in the region. Similar research involving the same logic was carried out at the same time by Gary Wright (1982) in an analysis of the population history of the Northwestern Plains of North America. Then, John Rick (1987) published a highly influential article titled "Dates as Data: An Examination of the Peruvian Preceramic Radiocarbon Record". Using a binned-dates approach, Rick argued that radiocarbon dates from Peru indicated important spatio-temporal population dynamics including changes in landscape use and settlement distributions throughout the preceramic period (roughly 20,000 to 3000 years ago). Importantly, Rick dedicated much of the paper to defending the notion of using radiocarbon dates as a proxy for population dynamics, spelling out potential biases and attempting to clarify the inferential logic behind the approach. The paper has since been cited hundreds of times and its title, "Dates as Data," is now a common label for the approach itself among archaeologists.

A key methodological development then occurred in the early 1990s. Up to this point, the methods for counting or summing radiocarbon-date densities involved uncalibrated dates. As Dye and Komori (1992) explained, however, aggregating only uncalibrated dates is a problem because it ignores the fact that environmental ratios of radiocarbon isotopes fluctuate through time. This fluctuation gives rise to the need to calibrate radiocarbon dates, which projects the date distribution from the radiocarbon timescale onto the calendar timescale (Taylor et al., 2014). Thus, Dye and Komori (1992) advocated applying the annual frequency distribution method developed by Law (Black and Green, 1977) to calibrated rather than uncalibrated dates. This was the first appearance of the now well-known "summed probability density function" (SPDF; see Figure 2), which is currently the most commonly-used proxy based on aggregated radiocarbon-dates.

The proxies have become extremely popular in the last decade as relatively large databases of radiocarbon dates have become available. A Web-of-Science search for the topic "summed radiocarbon" in archaeological and interdisciplinary palaeoenvironmental science journals returned over 100 articles published since 2010 and a similar Google Scholar search returned over 400 articles. In palaeoenvironmental research the approach has been widely adopted, and the proxies have been used to study a variety of target phenomena including past sea level changes, fire-regime surges, and climatic changes more generally (e.g. Bleicher, 2013; Mooney et al., 2011; Pierce et al., 2004; Thorndycraft and Benito, 2006). Archaeologists have been particularly enthusiastic, routinely using radiocarbon dates as a proxy for past population levels (e.g. Armit et al., 2013; Collard et al., 2010; Colledge et al., 2019; Faulkner, 2011; Gamble et al., 2005; Hannah and McLaughlin, 2019; Leipe et al., 2019; Lepofsky et al., 2005; Mclaughlin et al., 2018; Prentiss et al., 2014; Schulting, 2010; Shennan, 2013; Steele, 2010; Turney and Brown, 2007).

Since their introduction in the 1970s, however, scholars have also been aware of several key sources of bias affecting the proxies (e.g. Armit et al., 2013; Bamforth and Grund, 2012; Bleicher, 2013; Brown, 2015; Contreras and Meadows, 2014; Crema et al., 2017; Deacon, 1974; Kerr and McCormick, 2014; Manning and Timpson, 2014; Rick, 1987; Williams, 2012). These include radiocarbon sample quality (Brown, 2015), the relationship between a given sample and its sedimentary context (Geyh, 1969, 1971), spatio-temporal sampling sufficiency (Crema et al., 2017; Deacon, 1974; Rick, 1987; Williams, 2012), taphonomic processes (Brown, 2015; Surovell and Brantingham, 2007; Williams, 2012), calibration artifacts (Bamforth and Grund, 2012; Brown, 2015; Geyh, 1980; Kerr and McCormick, 2014), and – in the case of archaeological research – cultural, technological, and economic effects (Rick, 1987).

The widespread recognition of these biases has led to a number of methodological recommendations (see Crema et al., 2017; Williams, 2012, for more detailed reviews). To address issues of sample quality, several scholars have advocated carefully selecting radiocarbon dates based on the perceived reliability of the dates, quality of the dated material, and consideration of depositional context – together referred to as "chronometric hygiene"

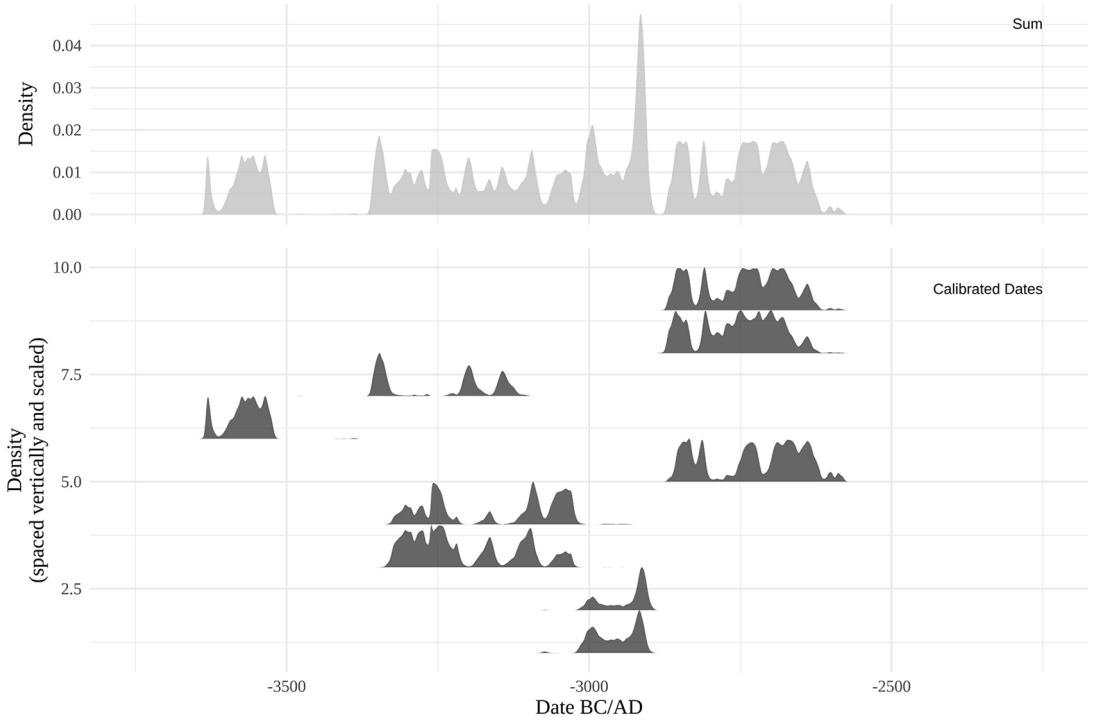


Figure 2. Example SPDF. Top panel (labelled "Sum") shows an SPDF based on the individual calibrated radiocarbon-date densities in the bottom panel (labelled "Calibrated Dates").

(e.g. Collard et al., 2010; Ebert et al., 2017; Hoggarth et al., 2016; Shennan, 2013). To overcome sampling issues, most scholars argue that large databases of dates are required (e.g. Shennan, 2013; Williams, 2012), although little agreement exists regarding what constitutes a good sample size in this setting. And, with respect to taphonomic bias, Surovell et al. (2009) suggested that independently dated tephra-based proxies for taphonomic loss could be used to correct the preservation bias in the proxies.

In addition, there have also been recent methodological advances concerning the proxies. These improvements can be divided into two groups. One focuses primarily on null hypothesis testing – comparing an empirical SPDF to one that acts as a benchmark for determining whether the empirical SPDF differs in a statistically significant way. For the approaches in this group, observed data are either compared to SPDFs of dates simulated from theoretical growth models (Crema and Kobayashi, 2020; Manning and Timpson, 2014; Shennan et al., 2013; Wicks and Mithen, 2014) or to a baseline distribution derived from the observed data with a permutation procedure (Crema et al., 2016, 2017). The former attempts to separate the target process from well-known spurious features introduced by the calibration process (e.g. Manning and Timpson, 2014; Shennan et al., 2013), while the latter attempts to account for spurious patterns produced by sampling variability (variation in spatial sampling intensity, specifically) (e.g. Crema and Kobayashi, 2020). In contrast, the other group focuses on improving the way date densities are summarised. This includes methods like sample bootstrapping (McLaughlin, 2019), Bayesian Gaussian mixture models (unpublished function in BChron, an R package for Bayesian radiocarbon date calibration; http://andrewcparnell.github.io/Bchron/), composite kernel density estimation (CKDE) (Brown, 2017), and partially-Bayesian kernel density estimation (KDE) (Bronk Ramsey, 2017). The density estimation and mixture model approaches limit the impact of calibration curve features resulting in smooth estimates. They also produce uncertainty envelopes that help distinguish potentially important variation from spurious fluctuations caused by the calibration curve (Bronk Ramsey, 2017).

The KDE-based approaches are perhaps the most significant developments because they involve a change in the fundamental way radiocarbon-dates are aggregated (Bronk Ramsey, 2017; Brown, 2017). KDE is a commonly used non-parametric method for estimating the continuous probability density of a random variable given a finite set of realizations of that variable (Silverman, 1986). In the case of proxies based on aggregated radiocarbon-dates, the realizations are an observed set of dates while the desired density is the aggregate temporal distribution of those dates. The kernel is essentially a moving window that computes a weighted-sum of the number of radiocarbon-dated events that occurred in a given time. It assigns weights based on the temporal distance from the center of the kernel to the date of a given sample. The closer a given date is to the center of the kernel, the more it contributes to the level of the proxy at the corresponding time. Of course, the dates are uncertain and that uncertainty is expressed by a radiocarbon-date distribution. So, both Bronk Ramsey (2017) and Brown (2017) have suggested algorithms to incorporate that uncertainty into a KDE. These algorithms sample the individual radiocarbon date distributions producing a set of probable dates, one for each event in a given database. Then, they apply a KDE to the random sample of event dates to produce a smooth estimate of temporal event density. The resulting sample "curve" is also referred to simply as a KDE. This process of sampling and density estimation is repeated a large number times yielding a set of KDEs. These KDEs are then combined to produce a single average KDE – which we will refer to as KDEa to distinguish it. The set of individual KDEs can then be used to produce the uncertainty envelope mentioned earlier.

Proxies based on aggregate radiocarbon-dates have been used so far in at least two distinct ways. The first way we call the "integral approach." Researchers employing this approach have used the proxies as approximately indicative of the total temporal distribution of events and their corresponding temporal uncertainties in a given database (Bronk Ramsey, 2017). Peaks or rises in the SPDF, for instance, are thought to indicate more events during the interval of time beneath them. This approach involves looking at the area under an SPDF and observing that more/less of that area is concentrated in one interval or another within the span of all of the relevant date densities (see Figure 3). The interpretation is downward-looking and horizontally oriented, focused on intervals of the time axis. In a sense, this view is analogous to the

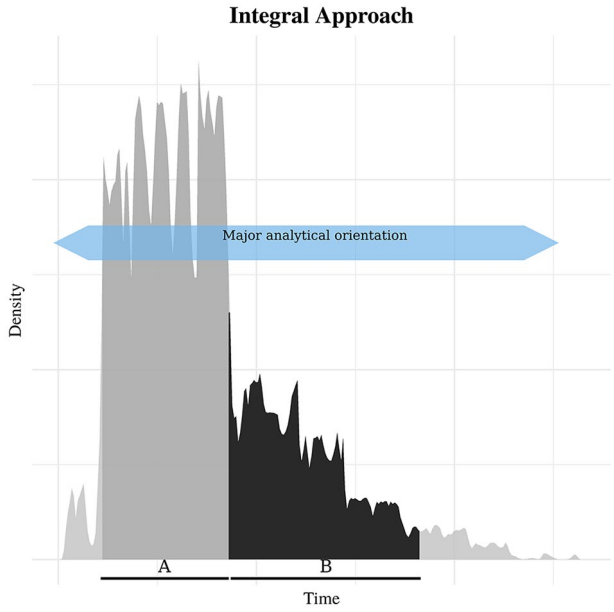


Figure 3. This graphic represents the "integral approach" with a simulated SPDF. Areas under the proxy are compared and interpreted like a probability density function. In this example, area A contains roughly 68% of the total area enclosed by the proxy, so it could be said that most of the total chronological information in the dataset – including uncertainty – is contained in area A, which is defined in part by a corresponding interval of time. Since it is larger than area B, one could argue that given the uncertainties in the underlying event times, most of the events have a higher likelihood of having occurred in interval A. The primary concern under the integral approach is the total temporal distribution of chronological information in the underlying dataset – that is, the distribution of dates for the events in question. The major analytical orientation, therefore, is horizontal, concerning the time-axis.

interpretation of an individual radiocarbon date density – that is, the probability that a given event dates to a given interval is proportional to the area under the density that spans the interval in question. In aggregate, then, more events are likely dated to periods over which the sum of many date densities has a larger area than the same sum over other periods.

A prominent example of the integral approach is a study by Shennan (2009) of population dynamics surrounding the appearance of the Linearbandkeramik (LBK) cultural phenomenon of Neolithic Western Europe. Shennan (2009) pointed to a set of SPDFs and claimed that "[l]ow Mesolithic population levels are succeeded by a massively increased LBK population." (p. 343) This claim appears to have been based on an integral interpretation of the relevant SPDF plots. The area under the SPDFs in those plots was much smaller in the interval before than after the accepted date for the onset of the LBK phenomenon. Thus, more of the dates – including their uncertainties – were located in the interval of time following the onset of the LBK. It follows that the events in question are more likely to be dated to the LBK period than before it. This pattern in turn, according to Shennan, indicates population levels must have increased dramatically *after* the onset of the LBK. The integral approach has been used extensively in numerous case studies (e.g. Becerra-Valdivia and Higham, 2020; Bishop, 2015; Boulanger and Lyman, 2013; Faulkner, 2011; Kerr et al., 2009; Lepofsky et al., 2005; Riede, 2008; Thorndycraft and Benito, 2006; Weninger et al., 2006).

We refer to the other way the proxies have been used as "point-wise" (or point-wise-like). This approach changes the orientation of analysis from the temporal distribution of events and their uncertainties (horizontal-looking) to through-time fluctuations in a given proxy (vertical-looking). It is a change from viewing the proxies as distributions of chronological information (Bronk Ramsey, 2017) to viewing them as if they are indicators of some process that may have changed through time (see Figure 4). The treatment of an SPDF/KDEa is necessarily point-wise whenever the level of the proxy at a given point in time needs to be compared to either its level at another time or another proxy at the same point in time. The proxy at time t is mapped directly to some other variable at time t. This type of comparison is necessary to make claims about covariation, estimate rates of change, or use regression models. Importantly, qualitative assessments can also be point-wise if an analyst is describing through-time changes explicitly and/or visually matching wiggles in a proxy with other variables. And, given small enough intervals of time, the previously described integral approach can become practically indistinguishable from a point-wise approach – that is, it can become point-wise-like – because changes in area under the curve start to become like simply changes in the level of the curve as the interval of time considered shrinks.

A recent example of point-wise comparisons is a paper on Late Quaternary Megafauna extinctions in North America by Broughton and Weitzel (2018). In that paper, the authors used linear regression to compare SPDFs representing megafauna populations to SPDFs representing human populations. By definition, the level of the megafauna population proxy at a given time was being mapped onto the level of the human population proxy at the same time. The average relationship between the pairings for every time under investigation was then represented by the estimated regression model parameters. The authors found a relationship between increasing human populations and declining megafauna populations for some species of megafauna between 15,000 BP and 10,000 BP. Quantitative point-wise comparisons like this occurred sporadically early on in the development of proxies based on aggregated radiocarbon-dates (e.g. Wendland and Bryson, 1974) and are becoming increasingly common (e.g. Bettinger, 2016; Ebert et al., 2017; Edinborough et al., 2017; Hannah and McLaughlin, 2019; Hinz et al., 2012; Plunkett et al., 2013; Robinson et al., 2019; Smith et al., 2008; Wang et al., 2014; Zahid et al., 2016).

Point-wise comparisons are important for understanding past processes, but it is not clear that they make sense where SPDF/KDEa proxies are involved. Ideally, point-wise comparisons would allow us to estimate rates of change in target processes and to identify potentially important casual forces by comparing a given proxy to the passage of time or to other proxy records, as we explained. Unfortunately, though, there is a good reason to think these comparisons may be unwise. Individual radiocarbon-date densities do not represent duration or through-time variation in a process that produces radiocarbon samples – they represent chronological uncertainty. It follows, then, that sums or aggregates of individual date densities also reflect chronological uncertainty in some way. Currently, to our knowledge, no attempts have been made to derive accurate point-wise interpretations for the established proxies, which leaves open crucial questions about how chronological uncertainty affects point-wise comparisons.

Here we attempt to rectify this situation. We first describe an attempt to derive interpretations for the SPDF and KDEa proxies and then we explore the downstream implications of those interpretations for point-wise comparisons. Lastly, we provide suggestions for future research that we think could improve dates-as-data approaches.

Interpreting proxies based on aggregated radiocarbon dates

For context, we imagined a hypothetical research scenario in which a large database of radiocarbon dates had been amassed. We further supposed that the individual dates represented a climatic or archaeological process. We also imagined that there was

no need for a complex Bayesian calibration model because the individual radiocarbon dates were not related stratigraphically, and we assumed a good representative sample. Lastly, we assumed that each event (e.g. archaeological or palaeoenvironmental deposit) was dated by precisely one radiocarbon date density – this meant that each event was dated by only one radiocarbon sample or that a given event was dated by a pooled density of multiple samples. These simplifications aided only in the exploration of the simplest kinds of SPDF/KDEa and more complex scenarios could be considered in the same manner.

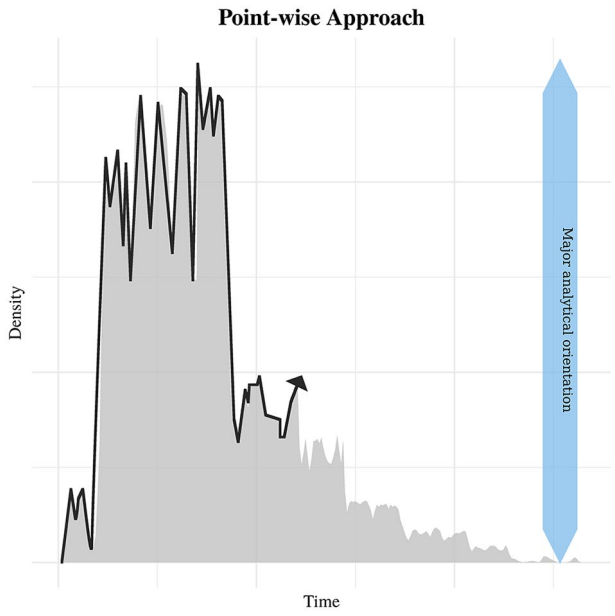


Figure 4. This graphic represents the "point-wise approach" with a simulated SPDF. Every point in the proxy is considered an observation that is indicative of the target process in some meaningful sense. This is required in order to estimate rates of change (i.e. compare the proxy to the passage of time itself) or to compare the proxy to other variables, like palaeoclimate reconstructions. The major analytical orientation, therefore, is vertical, concerning the measurement axis. That axis (the y-axis) is treated as if it measures the number of events dated to a particular time.

SPDF

First, we explored the SPDF. Let $p(\tau)$ be a single radiocarbon date density where τ refers to a single year in the domain (x-axis) of the density. The level of the density – that is, height with respect to the y-axis – at any given τ is proportional to the relative probability that the sample dates to τ compared to other times (see Figure 5) (Bronk Ramsey, 2009; Buck et al., 1996). Importantly, this interpretation holds for both calibrated and uncalibrated dates even though uncalibrated date densities – the kind returned from a radiocarbon dating lab – are used as "likelihoods" in Bayesian calibration models (Bronk Ramsey, 2009). Next, let $p_n(\tau)$ be a radiocarbon date density, n, from a set of N such densities. Since the domain (time) is common to all densities, τ in any $p_n(\tau)$ refers to a particular year in a given interval of years $[\tau_a, \tau_b]$. Then, the sum of the N densities for a given $\tau \in [\tau_a, \tau_b]$, denoted $S(\tau)$ is

$$S(\tau) = \sum_{n}^{N} p_n(\tau). \quad (1)$$

That $p_n(\tau)$ is proportional to the probability that event n occurred in year τ has implications for the interpretation of the summed function in equation (1). To see the implication clearly, we can expand the summation in equation (1) as follows:

$$S(\tau) = p_1(\tau) + p_2(\tau) + \ldots + p_N(\tau). \quad (2)$$

Since the terms in the sum are proportional to probabilities, the sum is a quantity proportional to the sum of the individual probabilities. We can, therefore, treat them as equivalent to probabilities and look to modern probability theory for the correct interpretation of $S(\tau)$.

Given standard probability theory, there are two potential interpretations of $S(\tau)$ (the SPDF). The difference between them depends on the relationship between individual events. Both interpretations involve the probability of the union of the individual events, which is calculated using the sum rule (Blitzstein and Hwang, 2015). The sum rule, as the name suggests, describes the situations in which probabilities can be summed (as in equation (2)) and the interpretations of the resulting sums.

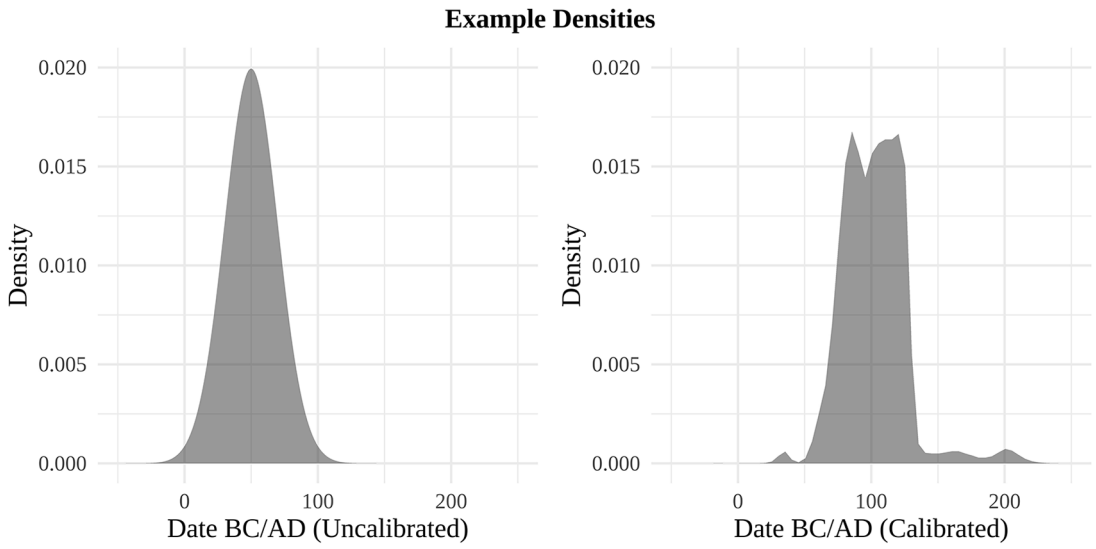


Figure 5. Example radiocarbon date densities. These densities represent an uncalibrated radiocarbon date of 1900 BP (50 CE) with an error of ±20 years. The plot on the left shows the uncalibrated distribution and the plot on the right shows the calibrated distribution produced by OxCal (Bronk Ramsey, 2009). The height of each curve for a given year indicates the probability that the sample dates to the relevant year compared to other years.

For the first interpretation, we assume that the events are *mutually exclusive*, meaning that they cannot co-occur. Consider, for example, events A and B, with corresponding probabilities, $P(A)$ and $P(B)$. When the events are mutually exclusive – that is, only one or the other can happen, not both – the sum rule is as follows (Blitzstein and Hwang, 2015),

$$P(A \cup B) = P(A) + P(B). \quad (3)$$

With respect to a pair of radiocarbon dates, mutual exclusivity would mean that the two events in question could not have happened in the same interval. Or, more formally, given two events, $p_1(\tau_i)$ and $p_2(\tau_j)$,

$$P(p_1(\tau_i) \cap p_2(\tau_j)) = 0 \quad | \quad i = j, \quad (4)$$

where τ_i and τ_j each refer to a date (interval, like a year or a decade) on a common time-axis.

Substituting into equation (3) two mutually exclusive radiocarbon-dated events – $p_i(\tau)$ and $p_j(\tau)$ – the right-hand side of equation (3) is equivalent to the SPDF in equation (2):

$$P(p_i(\tau) \cup p_j(\tau)) = S(\tau) = p_i(\tau) + p_j(\tau) \quad | \quad i \neq j. \quad (5)$$

The interpretation in this case is straightforward. It is the probability that *at least one* of the individual events occurred (Blitzstein and Hwang, 2015). In archaeological terms, it would be the probability that at least one of our imaginary radiocarbon-dated domestic buildings dates to a given time. This is essentially the interpretation described in the reference manual for OxCal under the "sum function" entry (https://c14.arch.ox.ac.uk/oxcalhelp/hlp_commands.html, accessed 2019-11-01).

If, however, the individual events are not mutually exclusive, the probability of the union must be calculated differently and this yields the second interpretation for $S(\tau)$. According to the sum rule, the probability of the union of non-mutually exclusive events is given by

$$P(A \cup B) = P(A) + P(B) - P(A \cap B), \quad (6)$$

where the last term – $P(A \cap B)$ – is the probability of the intersection of the events (Blitzstein and Hwang, 2015). This intersection refers to the portion of the probability space in which both events occur. In scientific terms, it would be the probability that both events (e.g. construction dates of domestic structures) date to the same time. But, the summation in eq. 1 contains no such term. Standard SPDFs are calculated without considering the probability of the intersection of events. Consequently, in the case of non-mutually-exclusive events, $S(\tau)$ is not proportional to the probability of the union of the individual events. Instead it must be greater by some quantity proportional to the probability of the intersection of the events in question. Again using two hypothetical radiocarbon-dated events, for example, equation (6) becomes,

$$P(p_i(\tau) \cup p_j(\tau)) = p_i(\tau) + p_j(\tau) - P(p_i(\tau) \cap p_j(\tau)) \quad | \quad i \neq j. \quad (7)$$

Thus, considering

$$S(\tau) = p_i(\tau) + p_j(\tau) \quad | \quad i \neq j, \quad (8)$$

it is clear that

$$S(\tau) > P(p_i(\tau) \cup p_j(\tau)) \quad | \quad i \neq j. \quad (9)$$

As a result, when the events in a given sample are not mutually exclusive, the level of an SPDF, $S(\tau)$, has no scientifically useful interpretation. Instead, SPDFs should be interpreted as indicating *through-time variations in some quantity greater than another quantity proportional to the probability that at least one of the events in the relevant data set occurred at a given time*. If that interpretation seems bewildering, then the point is clear.

The SPDF at any specific time conflates the number of events in a given database with uncertainty about their temporal positions. Even accounting for the probabilistic relationships among events – say by including the intersection term in equation (6) – the SPDF still refers to uncertainty. Its level at any given time indicates only the probability that at least one event occurred at the relevant time. Importantly, even the probability being indicated – at least one event – bears no necessary connection to the number of events that occurred. A few events with a high probability of occurrence at a given time would produce an SPDF value indistinguishable from a larger number of events that each had a low probability of having occurred at that time. Thus, variation from one time to the next in the level of the SPDF indicates only a change in relative probabilities, not a change in the number of events. There need not be any change in the number of underlying events for there to be fluctuations in level of an SPDF – and this would be true even in the absence of calibration artifacts or other biases.

Average KDE models

The KDEa models recently proposed by Bronk Ramsey (2017) and Brown (2017) are constructed in a very different way than a standard SPDF. Rather than summing radiocarbon-date densities at a given temporal resolution, the KDEa models are based on estimating the temporal density of radiocarbon samples. As a result, a KDEa model has a different interpretation than an SPDF. For our investigation, we primarily followed Bronk Ramsey's (2017) definitions and approach, but the same basic process underlies Brown's (2017) model.

Kernel density estimation is a method for approximating an unobservable continuous distribution from a finite observed sample (Silverman, 1986). In the context of KDEa models, the samples are a database of radiocarbon dates and the continuous distribution refers to the distribution of the relevant events in time. If there was no chronological uncertainty, the continuous density at any given time, τ, of a set of events, $\mathbf{t} = \{t_1, t_2, \ldots, t_n\}$, could be approximated by a simple kernel density estimator,

$$\check{f}(\tau) = \frac{1}{hn} \sum_{i=1}^{n} K\left(\frac{\tau - t_i}{h}\right), \quad (10)$$

where $K(\cdot)$ is a weighting function – usually Gaussian – that takes two inputs: (1) the distance from τ to each observation, and (2) the kernel bandwidth represented here by h (Bronk Ramsey, 2017). The kernel can be thought of as a kind of moving window and the bandwidth as the width of the window. As the window slides along the time-axis, it estimates the level of the density function at a time, τ, corresponding to its center. The density at the center of the kernel is a weighted sum of the number of events within its bandwidth, with weights inversely correlated to distance from the center to each event. But, unlike a moving window, the kernel density is continuous. So, instead of hard boundaries like the ones defining window edges, the kernel applies a smooth distance decay function. As a result, the weights applied to each event in the database are a function of distance and all of the events are included in the sum. The further away in time an observed event is from the center of the kernel, the less it counts toward the density at τ. The bandwidth parameter helps to determine how rapidly the assigned weights decay with distance – a

wider bandwidth would give more weight to observations further away in time.

In order to account for chronological uncertainty in the temporal locations of radiocarbon-dated events, the KDEa approaches involve a simulation. Essentially, the simulation has three main steps. First, a sample of potential dates, $\hat{\mathbf{t}} = \{\hat{t}_1, \hat{t}_2, \ldots, \hat{t}_n\}$, is randomly drawn from the set of radiocarbon-date densities. Importantly, the probability of a given date being drawn from a given density is proportional to the level of the density corresponding to that date.

Then, a bandwidth is determined. For Bronk Ramsey's approach, a modifier, g, is randomly drawn from a theoretical prior uniform distribution, $g \sim U(0,1)$, and the following kernel density function is used to estimate the density of the sample, $\hat{\mathbf{t}}$,

$$\hat{f}_{h_S}(\tau) = \frac{1}{g h_S n} \sum_{i=1}^{n} K\left(\frac{\tau - \hat{t}_i}{g h_S}\right). \quad (11)$$

where h_S refers to the commonly used Silverman (1986) bandwidth. Lastly, the likelihood of that density estimate and corresponding bandwidth modifier, g, is calculated. This process is repeated a large number of times to obtain a posterior density for g and a likelihood-weighted average KDEa model. The whole simulation is nested within the Markov-Chain Monte Carlo (MCMC) simulation used in OxCal (Bronk Ramsey, 2009) to calibrate radiocarbon dates. This means that prior information about stratigraphic relationships can be included and will be respected by the KDEa model. OxCal will also output an ensemble of KDE models, each of which represents a KDE calculated with a slightly different bandwidth because each will have used a slightly different g parameter. The ensemble can be used to provide a kind of uncertainty envelope for the primary, likelihood-weighted average KDEa model.

Brown's (2017) approach differs only in that it does not involve a Bayesian treatment of the bandwidth parameter. Instead, it uses a "plug-in" bandwidth that is determined by solving a separate equation (Jones et al., 1996) for every randomly drawn set of probable event dates. Ultimately, though, it also produces an average (composite) KDEa model and an ensemble based on randomly sampled dates from the set of densities.

We discovered a precise interpretation of both KDEa models by looking at the way they account for chronological uncertainty. As we explained, each iteration of the simulation used to produce a KDEa model involves randomly drawing a set of probable dates for a given set of radiocarbon samples. So, we let $\hat{\mathbf{t}} = \{\hat{t}_1, \hat{t}_2, \ldots, \hat{t}_n\}$ be the set of probable dates. Next we imagined focusing in on a specific time (τ) with a kernel that has a given bandwidth ($g h_S$). The level of the kernel function, K, at τ is primarily determined by the temporal distance between τ and a given set of dates, $\tau_{\Delta_i} = \tau - \hat{t}_i$. Treating the ratio outside the sum in eq. 11 as a proportionality constant and leaving it out we rewrote the equation more simply as follows,

$$\hat{f}(\tau) \propto \sum_{i=1}^{n} K\left(\frac{\tau_{\Delta_i}}{g h_S}\right). \quad (12)$$

Viewing this equation, it became clear that variation at τ between potential KDE models arises from repeatedly re-drawing $\hat{\mathbf{t}}$, which leads to variation in the numerator, τ_{Δ_i}. Greater distances between a given τ and $\hat{\mathbf{t}}$ cause $\hat{f}(\tau)$ to be lower for the relevant τ, while smaller distances cause it to be higher. Over the course of the simulation, this variability also leads to different optimal values for the bandwidth (or bandwidth modifier for Bronk Ramsey's KDEa approach). So during the simulation the primary source of variability in both the bandwidth and in $\hat{f}(\tau)$ is variability in τ_{Δ_i}. Since τ is just an arbitrary fixed position on the time-axis, the variability in τ_{Δ_i} also reflects variability in the temporal distance between radiocarbon samples around a given τ. Each time $\hat{\mathbf{t}}$ is drawn, the temporal distances between samples in $\hat{\mathbf{t}}$ change, which causes event-dates to cluster around some τ and disperse away from others. As the clustering fluctuates around a given τ, the level of the corresponding optimal KDEa model at that τ changes in response. It is these fluctuations in the shape of the optimal KDE that are captured in a KDE ensemble and combined in a likelihood-weighted average (the KDEa). Thus, we reasoned, the level of the KDEa model at a given time (τ) should be interpreted as *an estimate of the temporal density of radiocarbon samples weighted by uncertainty about the temporal distance between them*.

Importantly, the second part of this interpretation means that a KDEa model does not solely reflect through-time changes in the number of radiocarbon samples – it reflects chronological uncertainty as well. Just like the SPDF, changes in the level of the proxy are also a function of chronological uncertainty, not simply the number of radiocarbon samples dated to a particular time. This mixture of through-time variation in the number of dated events and chronological uncertainty is partly what makes the KDEa useful as a summary of chronological information for a large database of radiocarbon dates. It also, however, raises problems for interpreting specific fluctuations (point-wise differences) in the model. Ups and downs in the level of the curve cannot be directly interpreted as fluctuations in the number of events dated to a given time, nor the temporal density of events at a given time. The fluctuations also represent uncertainty about the temporal locations of the events in question, specifically uncertainty about the temporal distance between them. Consequently, point-wise fluctuations in the model cannot be straightforwardly interpreted as directly indicative of through-time fluctuations in the number of events or the corresponding target process. As Bronk Ramsey (2017) explained, the KDEa approach ". . . can be used to *summarize the distribution* of events. . ." in large ^{14}C databases (emphasis added) (p. 1831). It is a summary of chronological information, not a simple reflection of through-time variation in event-counts.

Discussion

Our investigation revealed a major underappreciated problem with proxies based on aggregated radiocarbon-dates: they do not represent the processes they are often thought to represent in a point-wise sense. It is important to stress that the problem does not extend to the method of radiocarbon dating more generally, only the use of aggregated/summed date densities as proxies for event counts. Both SPDF and KDEa proxies conflate chronological uncertainty with through-time variation in their target process. Neither, therefore, should be expected to clearly reflect through-time (point-wise) changes in any phenomenon related to the number of radiocarbon samples in palaeoenvironmental or archaeological records.

The conflation gives rise to several problems for point-wise analyses. We will explore the ones we think are the most obvious and important. Then, we will consider some of the main implications of these problems. Lastly, we will share some ideas for future research directions that we think could overcome them.

Analytical problems for point-wise approaches

Conceptual and statistical model mis-specification. Chiefly, the logic underpinning point-wise comparisons fails to hold because these proxies do not sufficiently indicate their intended target. This means that conceptual or quantitative models based on point-wise comparisons are mis-specified. Consequently, models

involving these proxies are almost certain to be misleading. While this is a problem for both quantitative and qualitative (visual) assessments, it is easiest to understand why the problem exists in quantitative terms.

Quantitatively, a point-wise comparison makes an explicit claim about how variables might be related. One variable, often called the "response" variable, is said to be a function of one or more other variables, often called "covariates" (Rencher and Schaalje, 2008). Referring back to Figure 4, the radiocarbon proxy would be the response variable and the covariate would be the date corresponding to a given proxy measurement. This type of point-wise relationship can be expressed by a simple equation. Imagine a single response variable, say an SPDF thought to represent the number of hearths dated to a given time. Also imagine a single covariate, say a palaeoclimate proxy for temperature, that we hypothesize might be related to hearth numbers. A simple equation describing the relationship between these variables could be written as,

$$Hearths_{time} = \beta Temp._{time}. \quad (13)$$

In this equation β – the regression coefficient – is a scaling factor or weight that relates the temperature at a given time to the number of hearths. The temperature at one point is equated with the number of hearths at the same point – a point-wise comparison. Within the context of statistical regression, it is also assumed that there is some uncertainty about the relationship. So, a term gets added to the right-hand side of the equation. This term changes the meaning of the equation from an exact relationship ("an increment of temperature will result in an exact proportional change in the number of hearths") to a probabilistic one whereby a change in temperature affects a change in the mean of a distribution of hearth-counts. The added term is often called an "error term" and conventionally denoted ϵ, but it also simply means the uncertainty around the mean response level. It determines the variance of the distribution of hearth-counts – how spread out observed hearth counts are around the mean level. So, equation (13) becomes,

$$Hearths_{time} = \beta Temp._{time} + \epsilon_{time}. \quad (14)$$

But, when an SPDF or KDEa model is used as a proxy for the response, the equation would actually look more like the following (with some abuse of notation for simplicity),

$$\begin{pmatrix} Hearths, \\ Dating\ Uncert. \end{pmatrix}_{time} = \beta Temp._{time} + \epsilon_{time}. \quad (15)$$

The revised model is saying that temperature determines the mean of some variable that is both the number of hearths *and* uncertainty about whether the hearths date to the relevant time. Information about dating uncertainty is being treated in the model as if it were information about hearth-count. We cannot tell the difference between an increase in hearth count and an increase in the relative likelihood that at least one hearth is dated to the relevant time. Consequently, the model cannot be used to say anything exclusively about hearth-count.

This conflation diminishes the model's utility as an explanation for through-time changes in the number of hearths. With the SPDF/KDEa proxies, it is as if someone poured two bags of marbles into a new bag, handed it to you, and then asked you to tell them how much each of the original bags weighed. Without more information you could never know. We cannot weigh the effect of temperature on mean hearth-count separate from its relationship to dating uncertainty. The marbles are all in the same bag.

This type of information misplacement is a kind of model mis-specification (Dennis et al., 2019; Rencher and Schaalje, 2008). The mathematical model is saying the wrong thing about the focal process. We want the model to tell us how temperature is related to the average number of hearths, but it is clearly telling us something else. Mis-specifications are known to produce biased and misleading results (Dennis et al., 2019), and the one we are describing here would have a similar effect. In fact, a recent simulation study by Brown (2017) demonstrated as much. Brown estimated a simple linear model that used a simulated SPDF as the response variable and found that the estimated regression coefficient – β in our example above – was severely biased. We also performed a simulation study showing the same thing (see Supplementary Information).

The mis-specification should trigger concern for a couple of reasons. One involves the biased estimates and misleading patterns we just discussed. It makes it much harder to meaningfully compare SPDF/KDEa proxies to other variables or to make comparisons involving the same proxy at different times. This problem precludes certain lines of research altogether – like identifying rates of change, or distinguishing the impacts of one potential causal factor from another.

The other reason is more grave. While some scholars may choose to see biases affecting statistical models as merely a technical nuisance, the main reason for concern is that the point-wise comparisons necessary for creating such a model make no scientific sense. These proxies mix up chronological uncertainty with their intended target in such a way that the two values cannot be unmixed. So, even if a covariate looks as if it may explain variation in the target process, what is actually being explained will always be a combination of variation in chronological uncertainty *and* the number of events. This is a serious scientific problem whether the point-wise comparison is formal and quantitative, or informal and visual.

Strange non-independence. A second, closely related, analytical problem involves observation independence. For both the SPDF and KDEa model, neighbouring observations have an unusual effect on each other. If one observation is proposed to be true – that is, accurately reflects the underlying number of events at a given time – the neighbouring observations *must* be false. This is because the events on which a given proxy is based can only have occurred at one time and not another. It is another consequence of the fact that individual date densities are expressions of uncertainty and, by extension, the aggregate proxies are too.

To illustrate, consider a single calibrated radiocarbon-date density. Variations in the level of the density indicate the relative probability that the event in question occurred at a given time. If it did occur in a particular time, however, it cannot have occurred in another time. The same logic applies to the SPDF and KDEa models. If we imagine for a moment that the level of an SPDF at time τ, for instance, reflected the number of events at time τ, the number of events at neighbouring times ($\tau \pm 1, \tau \pm 2, \ldots$) must necessarily be different than the proxy indicates. This is because events contributing to the level at time τ cannot simultaneously contribute to the level at neighbouring times. The problem is similar for the KDEa model – though, in that case, it is the magnitude of the contribution of a given event to the density at τ that must change in response to fixing any given event to a particular time.

This inter-temporal dependence has two immediately obvious consequences. One is that the error terms in a typical regression model are no longer independent. Referring back to the examples above (equations (13)–(15)), the dependence means that the ϵ_{time} term depends heavily on values at other times. For point-wise comparisons involving SPDF/KDEa proxies, this dependence cannot be easily accounted for with common time-series methods. For standard time-series models, the error structure is typically characterized by linear regression (Chatfield, 2004). Take, for example, moving average models. Such models account for

inter-temporal dependence in uncertainty by regressing the error term at one time on previous error terms with a linear model. This approach assumes that the error term at any given time is a linear combination of previous error terms weighted by regression coefficients. But, a database of radiocarbon dates has a much more complicated uncertainty structure. It is characterized by the joint-density of all the individual date densities and it would likely be far too complex to model linearly. That is not to say a model could not be developed, at least in theory, only that we are not aware of any established methods that could be used given the conflation of uncertainty and process in SPDF/KDEa proxies.

The other consequence of the unusually complex inter-temporal dependence in SPDF/KDEa proxies is that the proxies cannot be mapped to covariate observations in a stable, consistent way. It is impossible to determine whether a given proxy observation – say the level of an SPDF at time τ_2 – should be paired with a potential covariate observation at time τ_1, τ_2, τ_3, or some other time. Importantly, this problem is more than just additional measurement error, and as far as we can tell there is no standard statistical model that properly accounts for this strange non-independence where these proxies based on aggregated radiocarbon-dates are concerned.

Finite observations and infinite sample sizes. A third problem is that these proxies can give a false impression of the number of observations available with consequences for the identification of significant features/findings. In a classical statistical setting, "significance" often depends on the number of unique and independent pieces of information contributing to a perceived pattern. This is why standard text books devote space to discussing the importance of sample size (e.g. Devore, 2012; Ryan, 2013). Far from being unique to statistical applications, though, it is simply good scientific practice to be aware of how many observations are required to be confident that an identified pattern reflects reality.

The aggregate proxies, however, can be sampled at an arbitrarily high rate. This allows for a kind of quantitative alchemy whereby a finite number of individual radiocarbon-dated events can be transmuted into a sample of unlimited size. An individual radiocarbon-date density, for instance, is an approximation of an underlying function that can be produced at any resolution. Often the default for calibration programs is 5 or 10 years, but there is no hard cap in theory with the only limiting factor being a trade-off between spuriously high precision and accuracy. So, an SPDF or KDEa model could, for example, be annually resolved and span millennia. This would result in an apparent sample size in the multiple thousands, even if the SPDF was ultimately based on only a dozen dates.

A statistical model, unfortunately, could not "know" the difference – they are just equations, after all, and it is up to the one using them to decide whether they apply in a given case. Most common approaches (regressions or time-series methods, for example) entail counting the number of observations in the SPDF in order to determine sample size. This would mean that the sample size was the number of time-points at which the proxy was sampled, which is determined by its resolution, not the number of dates used to create the proxy in the first place. Scientifically, though, it is the number of dates that indicate how many unique pieces of information are involved in the analysis, at least with respect to models seeking to explain event count variation.

For point-wise comparisons, we think this sample size problem is unavoidable. While several archaeologists have cautioned that sample sizes need to be large for SPDFs/KDEas to be meaningful (e.g. Bishop, 2015; Contreras and Meadows, 2014; Williams, 2012), it is nonetheless always possible to oversample these models. Thus, these proxies will tend to give the impression that the sample size is much larger than it actually is. More importantly, there is no obvious relationship between the number of dates used to create a proxy and the appropriate sampling rate that should be used in order to avoid oversampling. Ultimately, the problem again stems from confusing chronological uncertainty with event counts. Single date densities, representing only one sample, are treated like representations of through-time processes when they are aggregated with the densities of other samples. Individually, the densities do reflect some latent, smooth process that has to do with radiocarbon-dating uncertainties (e.g. isotope instrumentation error, calibration curve uncertainties, through-time fluctuations in environmental carbon isotope ratios, etc.). But aggregated together and used as a proxy for a radiocarbon sample producing process, the smooth line is thought to reflect event-count. In the former case, one could justifiably assume that there is some smooth underlying process represented by a date density and then sample it at whatever rate was reasonable – this is the basis for Functional Data Analysis where a smooth estimate based on a finite sample actually does represent an underlying continuous process (Ramsay et al., 2009). With proxies based on aggregated radiocarbon-dates this is simply not the case. At least, the continuous underlying process that could be sampled is a reflection of chronological uncertainty, not through-time changes in the number of events. It is the latter most scholars are interested in and the way that the established proxies are often interpreted.

Misleading density-like structure. The last major analytical problem is that chronological uncertainty imposes a characteristic structure on these proxies that can be misleading. Chronological uncertainty is inherently density-like, by which we mean shaped like a density function (e.g. a bell-curve). Uncertainty about the timing of an event means that we assign some likelihood to the event having occurred at a given time and that likelihood declines with increasing temporal distance from that most likely time. As one might expect, this density-like structure applies to individual radiocarbon-date densities, sums of radiocarbon-date densities, and of course KDEas based on radiocarbon-date densities.

A density-like structure creates a confounding problem for point-wise approaches. Any density function has to have a positive finite integral – that is, the area between a continuous density function (curve) and its domain (x-axis) has to be entirely above the domain (positive everywhere) and it has to have a finite limit (the area cannot be infinitely large) (Blitzstein and Hwang, 2015). In more concrete terms, our uncertainty about the timing of one or more events is not infinite, so a curve representing that uncertainty must occupy a finite amount of time. This means that the curve can go up and down in any manner, but it must "begin" by going up and "end" at some point by going down. The same is true of aggregated density functions representing groups of individual events, like the SPDF and KDEa. Sometimes, a given proxy may exhibit several up/down fluctuations, but overall the up-then-down pattern will be consistent irrespective of the true through-time structure of the target process. Consequently, whatever the true relationship looks like for a given process–covariate pair, the relationship between a corresponding proxy and the covariate will be distorted. Ultimately, the distortion occurs because we are uncertain about the timing of the individual events in question and our uncertainty has a natural density-like structure. Importantly, this structure cannot be accounted for separately from variation in event-count when SPDF/KDEa models are used.

Implications

The problems we identified can lead scientists severely astray in point-wise analyses. Determining the magnitude of the problem depends on many factors that will be particular to a given dataset and analysis. These include the number of dates involved, the span of time under analysis, the level of calibration curve uncertainty, the slope of the calibration curve, and the variation in the

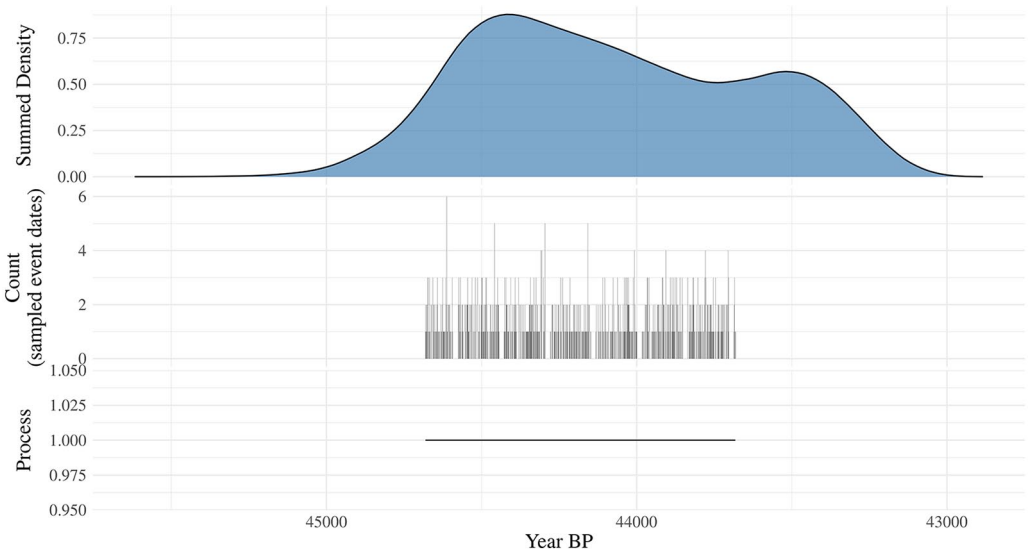


Figure 6. Simulated SPDF data. The bottom time series represents the target uniform process. The center series represents the count series corresponding to the random event-date sample of 1000 observations. The top series is the corresponding SPDF proxy.

target process. These factors will affect the "signal-to-noise" ratio – that is, the magnitude of variation in the target process relative to the level of chronological uncertainty. If, for example, the underlying target process changes dramatically over a given interval of interest, then certain features of an SPDF/KDEa proxy of that process may be vaguely representative assuming (1) a high enough sampling rate (number of dates per unit of time) and (2) low chronological uncertainty per date relative to the length of interval of interest. But, the chronological uncertainty will be present nonetheless and the problems we described above will continue to hinder attempts to recover underlying patterns. Importantly, the problem of inter-temporal dependence will still be present, and the overall density-like structure of the proxy will distort the true target process. Thus, faulty inferences are very likely, even when attempting to address an elementary research question like whether a given target process increased, decreased, or remained constant through time.

To demonstrate, we conducted a simple simulation. The target process in our simulation was a simple uniform function – meaning, no change through time and an average slope of zero. It would be analogous to, say, population levels that were constant through time. We then used standard regression models to try and estimate the slope value. The goal was to determine whether false-positive findings were likely and, if they were, to figure out whether the conflation between chronological uncertainty and process variation was likely to blame.

First, we created a uniform process over a 1000-year interval with a start date chosen at random from between 48,000 and 2000 BP (see Figure 6). We then sampled the uniform process randomly, drawing 1000 event-dates from it. Next, we used the "calBP.14C" function from the R package, "clam" (Blaauw, 2020), to derive uncalibrated dates from the event-date sample, which we subsequently calibrated with the "calibrate" function from the same package. We then created three time-series. One was an SPDF based on the simulated calibrated dates. The second was a count time series of the event-date sample – this series represents a high-fidelity sample of the true process with no chronological error. The last was a density estimate of the event-date sample, produced with R's built-in "density()" function. This series was approximately like an SPDF/KDEa model, but without any chronological error. It allowed us to isolate the distorting effects that smooth density estimates would be expected to have on regression models from the specific effect of conflating chronological uncertainty with process variation as the SPDF/KDEa proxies do.

With these data in hand, we ran three regressions using R's "glm()" function. In one, we used a Gamma-distributed model with the simulated SPDF as the response variable and time as the only covariate. For the second one we used the count-series as the response variable and time as the covariate in a Poisson model. For the last regression we produced another Gamma-distributed model with the chronological-error-free density estimate as the response variable.

The regression results were very clear. Unsurprisingly, the models involving the chronological-error-free time-series indicated that the slope of the target process was indistinguishable from zero (see Figure 7). The SPDF regression, however, produced a very different result. The target estimate was severely biased, with the mean very far from zero (see Figure 7). Crucially, it also indicated that the non-zero effect was highly statistically significant, well beyond the standard 95% or 99% confidence levels. We re-ran this simulation repeatedly with consistent results. Using an SPDF in a point-wise analysis can be very misleading for the reasons we outlined. The simulation described in our SI demonstrates similar problems for the KDEa proxy.

These problems have important implications for past and future research. A major implication for previous research is that published findings based on SPDF/KDEa proxies could be misleading. In more qualitative studies involving no regression models (e.g. Bradtmöller et al., 2012), narratives about through-time processes like demographic changes or fluctuations in sea level may be telling the wrong story. In more quantitative research involving correlation and/or regression (e.g. Bettinger, 2016), hypothesis tests could be invalid and parameter estimates false or biased, calling into question inferences drawn from such findings. In light of the potential problems we have identified, this body of research should be viewed with some measure of suspicion, in our view.

A major implication for future research is that these proxies should be avoided when point-wise interpretations are necessary or implied. They could instead be profitably used for addressing some questions about chronological relationships. Research into, for example, whether a collection of events pre- or post-dates a fixed date could also be based on the established proxies (e.g. Bronk Ramsey, 2017). This is the "integral approach" we described earlier whereby the analysis is focused on the relative sizes of areas under the proxies and the corresponding intervals of time covered. Comparing these areas does not appear to be conflating process variation and chronological uncertainty in the same manner as the "point-wise" approach does. It is important to

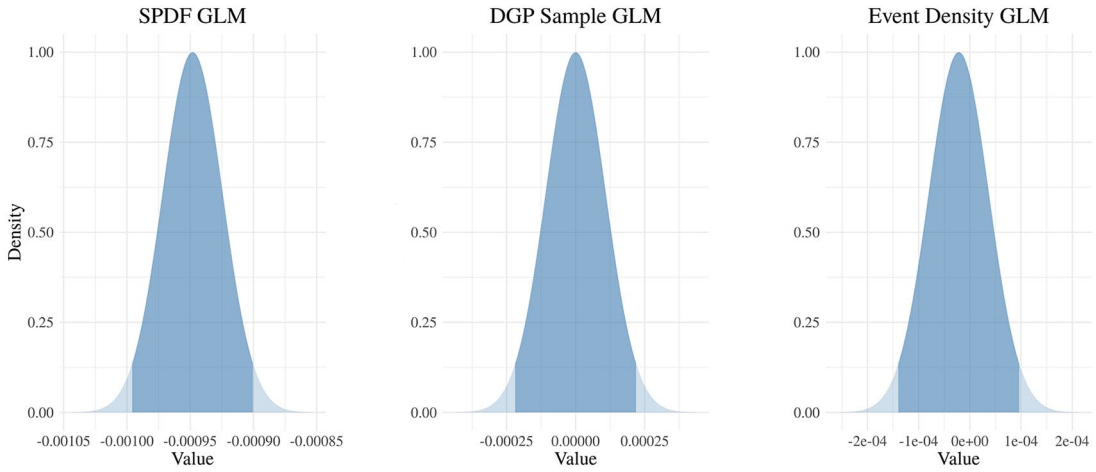


Figure 7. Regression simulation parameter estimates. The two densities were created by using the regression coefficient parameters estimated from our two regression models. The density on the left represents the sampling distribution for the regression coefficient related to time in the SPDF regression model. The density in the center represents the sampling distribution for the corresponding coefficient in the Poisson model based on the sample of event-dates. And, the distribution on the left represents the same information for the model involving the density estimate of the event-dates. The dark blue areas indicate the 95% confidence regions and the lighter blue areas represent the 99% regions.

emphasize, though, that more systematic and mathematically grounded approaches exist for estimating the relevant probabilities (Bronk Ramsey, 2009; Buck et al., 1996) and, at best, an SPDF or KDEa may only be appropriate for visualization. Research intended to improve our understanding of through-time fluctuations in the number of radiocarbon-dated events, however, should not be based on these proxies.

Directions for future work

We imagine future research proceeding in two main directions. One involves a review of previous studies. Many of the published studies involving proxies based on aggregated radiocarbon-dates should be re-evaluated to determine whether the problems we identified undermine previous conclusions. This is especially the case where interpretations depend on the point-wise approach. We recognize that determining whether a given argument or study depends on point-wise interpretations may be challenging for qualitative cases because of the vagueness of the language sometimes used. But, reports that highlight short-term fluctuations in these proxies, or reports that involve claims about rates of change or magnitude of fluctuations, should probably be considered suspect. Similarly, claims about "events," like rapid declines in event counts, must be relying on point-wise comparisons because identifying a "rapid change" requires one to evaluate the difference between levels of a given proxy at neighbouring or nearby times. Quantitative research, in contrast, will be easier to evaluate in this regard because it is necessarily point-wise and the claims being made must have been explicitly spelled out.

The other direction for future research involves methodological development. As is hopefully clear, we think that the main problem with using dates-as-data in point-wise analyses is the confounding between chronological uncertainty and process variation. There are potentially useful alternative approaches, though. In a study on chronological uncertainty in layer-counted archives, for example, Boers et al. (2017) proposed a method of re-projecting uncertainty from the time-domain onto the measurement domain. It takes chronological uncertainty and turns it into measurement uncertainty, which would make point-wise comparisons more sensible. Rather than conflating chronological uncertainty with event counts, the reprojected proxy would ideally indicate the probable number of events that occurred in a given time. That way, point-wise conceptual or quantitative models would be explaining variation in one dimension exclusively. The method they proposed was developed for use with layer-counted archives – e.g. ice cores, or varved lake sediments – but it might be adapted for use with radiocarbon-dated events. A persistent challenge will be finding a way to correctly account for the non-independence between observations, but further research is warranted.

Another recent development involves a Bayesian regression approach for analyzing radiocarbon-dated event-count series. These Radiocarbon-dated Event Count (REC) models have been tested with simulated data and appear to be able to separate target process dynamics (event-counts) from chronological uncertainty (Carleton, 2020). REC models effectively extend the basis of the KDEa approaches (Bronk Ramsey, 2017; Brown, 2017) but are directed at regression model estimation rather than density estimation. They are based on ensembles of potential event-count sequences created by sampling the corresponding calibrated radiocarbon-date densities – a Radiocarbon Event-Count Ensemble, or RECE. Each member of the RECE is then placed in a suitable regression model. The individual models are nested in a multi-level Bayesian framework so that priors for target parameters (e.g. regression coefficients) can be specified. As such, no individual regression model conflates uncertainty with process variation – these types of information are kept separate. At the same time, chronological uncertainty is accounted for in the posterior distributions of the top-level parameters. Further methodological exploration of REC models could go some way toward making the dates-as-data paradigm viable in the context of point-wise comparisons aimed at rate estimation, model comparison, and standard hypothesis testing.

Conclusion

Proxies for radiocarbon-dated event counts are tempting to use and have appeared in hundreds of scholarly articles since their inception. They are generally intended to represent through-time fluctuations in the amount of radiocarbon in the archaeological and palaeoenvironmental records. These through-time fluctuations are thought to be caused by variation in one or more radiocarbon-producing target processes, including past human activity, population level changes, sea-level changes, and surging fire regimes. However, there are lurking problems that make these proxies unsuitable representations of their targets when viewed in a point-wise way. The main problem is that these proxies do not clearly reflect

through-time variation in the number of radiocarbon-dated events in a given database. Rather, they combine through-time variation with chronological uncertainty. Unfortunately, this conflation is easy to miss, leading to biases and faulty interpretations. We, therefore, urge scholars to think carefully about how the proxies are being interpreted and whether they are appropriate for a given research agenda. Any researchers intent on using the proxies will need to explain precisely why the problems we have identified are not materially affecting their interpretations. Alternative approaches that avoid the conflation are under development and scholars interested in using radiocarbon-dates as data should consider using those methods instead.

Acknowledgements

We would like to thank the editor, Professor John Matthews, and three anonymous reviewers for their helpful comments and feedback. We would also like to thank Professor Caitlin Buck for her helpful correspondence early on in the research process.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Funding for this research was provided by the Max Planck Society.

ORCID iDs

W. Christopher Carleton https://orcid.org/0000-0001-7463-8638

Huw S. Groucutt https://orcid.org/0000-0002-9111-1720

Supplemental material

Supplemental material for this article is available online.

References

Armit I, Swindles GT and Becker K (2013) From dates to demography in later prehistoric Ireland? Experimental approaches to the meta-analysis of large 14C data-sets. *Journal of Archaeological Science* 40(1): 433–438.

Bamforth DB and Grund B (2012) Radiocarbon calibration curves, summed probability distributions, and early Paleoindian population trends in North America. *Journal of Archaeological Science* 39(6): 1768–1774.

Becerra-Valdivia L and Higham T (2020) The timing and effect of the earliest human arrivals in North America. *Nature* 584: 93–97.

Berry M (1982) *Time, Space, and Transition in Anasazi Prehistory*. Salt Lake City, Utah: University of Utah Press.

Bettinger RL (2016) Prehistoric hunter-gatherer population growth rates rival those of agriculturalists 113(4): 812–814.

Bishop RR (2015) Did Late Neolithic farming fail or flourish? A Scottish perspective on the evidence for Late Neolithic arable cultivation in the British Isles. *World Archaeology* 47(5): 834–855.

Blaauw M (2020) *clam: Classical Age-Depth Modelling of Cores from Deposits*. R package version 2.3.4.

Black S and Green RC (1977) Radiocarbon dates from the Solomon Islands to 1975. *Oceanic Prehistory Records* 4: 1–56.

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

Blitzstein JK and Hwang J (2015) *Introduction to Probability: Texts in Statistical Science*. Boca Raton, FL: CRC Press.

Boers N, Goswami B and Ghil M (2017) A complete representation of uncertainties in layer-counted paleoclimatic archives. *Climate of the Past* 13(9): 1169–1180.

Boulanger MT and Lyman RL (2013) Northeastern North American Pleistocene megafauna chronologically overlapped minimally with Paleoindians. *Quaternary Science Reviews* 85: 35–46.

Bradtmöller M, Pastoors A, Weninger B et al. (2012) The repeated replacement model Rapid climate change and population dynamics in Late Pleistocene Europe. *Quaternary International* 247: 38–49.

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

Bronk Ramsey C (2017) Methods for summarizing radiocarbon datasets. *Radiocarbon* 59(6): 1809–1833.

Broughton JM and Weitzel EM (2018) Population reconstructions for humans and megafauna suggest mixed causes for North American Pleistocene extinctions. *Nature Communications* 9(1): 1–12.

Brown WA (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.

Brown WA (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.

Buck CE, Cavanagh WG and Litton CD (1996) *Bayesian Approach to Interpreting Archaeological Data*. Chichester, UK: John Wiley and Sons, Inc.

Carleton W (2020) Evaluating bayesian radiocarbon-dated event count (REC) models for the study of long-term human and environmental processes. *Journal of Quaternary Science*. Epub ahead of print 20 October 2020. DOI: 10.1002/jqs.3256.

Chatfield C (2004) *The Analysis of Time Series: An Introduction*, 6th edn. Boca Raton, FL: Chapman & Hall/CRC.

Collard M, Buchanan B, Hamilton MJ et al. (2010) Spatiotemporal dynamics of the Clovis-Folsom transition. *Journal of Archaeological Science* 37(10): 2513–2519.

Colledge S, Conolly J, Crema E et al. (2019) Neolithic population crash in northwest Europe associated with agricultural crisis. *Quaternary Research* 92(3): 686–707.

Contreras DA and 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.

Crema ER, Bevan A and Shennan S (2017) Spatio-temporal approaches to archaeological radiocarbon dates. *Journal of Archaeological Science* 87: 1–9.

Crema ER, Habu J, Kobayashi K et al. (2016) Summed probability distribution of 14C dates suggests regional divergences in the population dynamics of the Jomon period in Eastern Japan. *PLoS ONE* 11(4): e0154809.

Crema ER and 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. DOI:10.1016/j.jas.2020.105136.

d'Alpoim Guedes JA, Crabtree SA, Bocinsky RK et al. (2016) Twenty-first century approaches to ancient problems: Climate and society. *Proceedings of the National Academy of Sciences* 113(51): 14483–14491.

Deacon J (1974) Patterning in the radiocarbon dates for the Wilton/Smithfield complex in Southern Africa. *The South African Archaeological Bulletin* 29(113): 3–18.

Dennis B, Ponciano JM, Taper ML et al. (2019) Errors in statistical inference under model misspecification: Evidence, hypothesis testing, and AIC. *Frontiers in Ecology and Evolution* 7: 372.

Devore JL (2012) *Probability and Statistics for Engineering and the Sciences*, 8th edn. Boston, MA: Brooks/Cole, Cengage Learning.

Dye T and Komori E (1992) Computer programs for creating cumulative probability curves and annual frequency distribution diagrams with radiocarbon dates. *New Zealand Journal of Archaeology* 14: 35–43.

Ebert CE, Peniche May N, Culleton BJ et al. (2017) Regional response to drought during the formation and decline of Preclassic Maya societies. *Quaternary Science Reviews* 173: 211–235.

Edinborough K, Porčić M, Martindale A et al. (2017) Radiocarbon test for demographic events in written and oral history. *Proceedings of the National Academy of Sciences* 114(47): 12436–12441.

Faulkner P (2011) Late-Holocene mollusc exploitation and changing near-shore environments: A case study from the coastal margin of Blue Mud Bay, northern Australia. *Environmental Archaeology* 16(2): 173–180.

Gamble C, Davies W, Pettitt P et al. (2005) The archaeological and genetic foundations of the European population during the Late Glacial: Implications for "agricultural thinking." *Cambridge Archaeological Journal* 15(2): 193–223.

Geyh M (1969) Versuch einer chronologischen Gliederung des marinen Holozäns an der Nordseeküste mit Hilfe der statistischen Auswertung von 14C-Daten. *Zeitschrift der Deutschen Geologischen Gesellschaft* 118(2): 351–360.

Geyh MA (1971) Middle and young Holocene sea-level changes as global contemporary events. *Geologiska Föreningen i Stockholm Förhandlingar* 93(4): 679–692.

Geyh MA (1980) Holocene sea-level history: Case study of the statistical evaluation of ^{14}C dates. *Radiocarbon* 22(3): 695–704.

Hannah E and McLaughlin R (2019) Long-term archaeological perspectives on new genomic and environmental evidence from early medieval Ireland. *Journal of Archaeological Science* 106: 23–28.

Hinz M, Feeser I, Sjögren KGG et al. (2012) Demography and the intensity of cultural activities: An evaluation of Funnel Beaker Societies (4200–2800 cal BC). *Journal of Archaeological Science* 39(10): 3331–3340.

Hoggarth JA, Breitenbach SF, Culleton BJ et al. (2016) The political collapse of Chichén Itzá in climatic and cultural context. *Global and Planetary Change* 138: 25–42.

Jones MC, Marron JS and Sheather SJ (1996) A brief survey of bandwidth selection for density estimation. *Journal of the American statistical association* 91(433): 401–407.

Kerr TR and McCormick F (2014) Statistics, sunspots and settlement: Influences on sum of probability curves. *Journal of Archaeological Science* 41: 493–501.

Kerr TR, Swindles GT and Plunkett G (2009) Making hay while the sun shines? Socio-economic change, cereal production and climatic deterioration in Early Medieval Ireland. *Journal of Archaeological Science* 36(12): 2868–2874.

Leipe C, Long T, Sergusheva EA et al. (2019) Discontinuous spread of millet agriculture in eastern Asia and prehistoric population dynamics. *Science Advances* 5(9): eaax6225.

Lepofsky D, Lertzman K, Hallett D et al. (2005) Climate change and culture change on the southern coast of British Columbia 2400-1200 cal. B.P.: An hypothesis. *American Antiquity* 70(2): 267–293.

Mclaughlin R, Hannah E and Coyle-McClung L (2018) Frequency analyses of historical and archaeological datasets reveal the same pattern of declining sociocultural activity in 9th to 10th century CE Ireland. *Cliodynamics: The Journal of Quantitative History and Cultural Evolution* 9(1): 1–24.

McLaughlin TR (2019) On applications of spacetime modelling with open-source ^{14}C age calibration. *Journal of Archaeological Method and Theory* 26(2): 479–501.

Manning K and Timpson A (2014) The demographic response to holocene climate change in the Sahara. *Quaternary Science Reviews* 101: 28–35.

Mooney SD, Harrison SP, Bartlein PJ et al. (2011) Late Quaternary fire regimes of Australasia. *Quaternary Science Reviews* 30(1–2): 28–46.

Pierce JL, Meyer GA and Timothy Jull AJ (2004) Fire-induced erosion and millennial-scale climate change in northern ponderosa pine forests. *Nature* 432(7013): 87–90.

Plunkett G, McDermott C, Swindles GT et al. (2013) Environmental indifference? A critique of environmentally deterministic theories of peatland archaeological site construction in Ireland. *Quaternary Science Reviews* 61: 17–31.

Prentiss AM, Cail HS and Smith LM (2014) At the Malthusian ceiling: Subsistence and inequality at Bridge River, British Columbia. *Journal of Anthropological Archaeology* 33(1): 34–48.

Ramsay J, Hooker G and Graves S (2009) *Functional Data Analysis with R and MATLAB*. New York, NY: Springer.

Rencher AC and Schaalje GB (2008) *Linear Models in Statistics*. Hoboken, NJ: John Wiley & Sons, Inc.

Rick JW (1987) Dates as data: An examination of the Peruvian preceramic radiocarbon record. *American Antiquity* 52(1): 55–73.

Riede F (2008) The Laacher See-eruption (12,920 BP) and material culture change at the end of the Allerød in Northern Europe. *Journal of Archaeological Science* 35(3): 591–599.

Riede F (2009) Climate and demography in early prehistory: Using calibrated (14)C dates as population proxies. *Human Biology* 81(2–3): 309–337.

Robinson E, Zahid HJ, Codding BF et al. (2019) Spatiotemporal dynamics of prehistoric human population growth: Radiocarbon "dates as data" and population ecology models. *Journal of Archaeological Science* 101: 63–71.

Ryan TP (2013) *Sample Size Determination and Power*. Hoboken, NJ: John Wiley & Sons, Inc.

Schulting R (2010) Holocene environmental change and the Mesolithic-Neolithic transition in north-west Europe: Revisiting two models. *Environmental Archaeology* 15(2): 160–172.

Shennan S (2009) Evolutionary demography and the population history of the European early neolithic. *Human Biology* 81(2–3): 339–355.

Shennan S (2013) Demographic continuities and discontinuities in Neolithic Europe: Evidence, methods and implications. *Journal of Archaeological Method and Theory* 20(2): 300–311.

Shennan S, Downey SS, Timpson A et al. (2013) Regional population collapse followed initial agriculture booms in mid-Holocene Europe. *Nature Communications* 4(1): 2486.

Silverman BW (1986) *Density Estimation for Statistics and Data Analysis*. London: Chapman and Hall.

Smith M, Williams A, Turney C et al. (2008) Humanenvironment interactions in Australian drylands: Exploratory time-series analysis of archaeological records. *The Holocene* 18(3): 389–401.

Steele J (2010) Radiocarbon dates as data: Quantitative strategies for estimating colonization front speeds and event densities. *Journal of Archaeological Science* 37(8): 2017–2030.

Surovell TA and Brantingham PJ (2007) A note on the use of temporal frequency distributions in studies of prehistoric demography. *Journal of Archaeological Science* 34(11): 1868–1877.

Surovell TA, Byrd Finley J, Smith GM et al. (2009) Correcting temporal frequency distributions for taphonomic bias. *Journal of Archaeological Science* 36(8): 1715–1724.

Taylor RE, Bar-Yosef O and Renfrew C (2014) *Radiocarbon Dating: An Archaeological Perspective*, 2nd edn. Albuquerque, New Mexico: Routledge.

Thorndycraft VR and Benito G (2006) The Holocene fluvial chronology of Spain: Evidence from a newly compiled radiocarbon database. *Quaternary Science Reviews* 25(3–4): 223–234.

Turney CS and Brown H (2007) Catastrophic early Holocene sea level rise, human migration and the Neolithic transition in Europe. *Quaternary Science Reviews* 26(17–18): 2036–2041.

Wang C, Lu H, Zhang J et al. (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.

Wendland WM and Bryson RA (1974) Dating climatic episodes of the Holocene. *Quaternary Research* 4(1): 9–24.

Weninger B, Alram-Stern E, Bauer E et al. (2006) Climate forcing due to the 8200cal yr BP event observed at Early Neolithic sites in the eastern Mediterranean. *Quaternary Research* 66(3): 401–420.

Wicks K and Mithen S (2014) The impact of the abrupt 8.2ka cold event on the Mesolithic population of western Scotland: A Bayesian analysis using 'activity events' as a population proxy. *Journal of Archaeological Sciences* 45(1): 240–269.

Williams AN (2012) The use of summed radiocarbon probability distributions in archaeology: A review of methods. *Journal of Archaeological Science* 39(3): 578–589.

Wright GA (1982) Notes on chronological problems on the northwestern plains and adjacent country. *Plains Anthropologist* 27: 145–160.

Zahid HJ, Robinson E and Kelly RL (2016) Agriculture, population growth, and statistical analysis of the radiocarbon record. *Proceedings of the National Academy of Sciences of the United States of America* 113(4): 931–935.