

Evaluating Bayesian Radiocarbon-dated Event Count (REC) models for the study of long-term human and environmental processes

W. CHRISTOPHER CARLETON* 

Extreme Events Research Group, Max Planck Institutes for Chemical Ecology, The Science of Human History, and Biogeochemistry, Jena, Germany

Received 8 April 2020; Revised 22 September 2020; Accepted 24 September 2020

ABSTRACT: Chronological uncertainty complicates attempts to use radiocarbon dates as proxies for processes such as human population growth/decline, forest fires and marine ingression. Established approaches involve turning databases of radiocarbon-date densities into single summary proxies that cannot fully account for chronological uncertainty. Here, I use simulated data to explore an alternative Bayesian approach that instead models the data as what they are, namely radiocarbon-dated event counts. The approach involves assessing possible event-count sequences by sampling radiocarbon date densities and then applying a Markov Chain Monte Carlo method to estimate the parameters of an appropriate count-based regression model. The regressions based on individual sampled sequences were placed in a multilevel framework, which allowed for the estimation of hyperparameters that account for chronological uncertainty in individual event times. Two processes were used to produce simulated data. One represented a simple monotonic change in event-counts and the other was based on a real palaeoclimate proxy record. In both cases, the method produced estimates that had the correct sign and were consistently biased towards zero. These results indicate that the approach is widely applicable and could form the basis of a new class of quantitative models for use in exploring long-term human and environmental processes.

© 2020 The Authors. *Journal of Quaternary Science* Published by John Wiley & Sons Ltd.

KEYWORDS: archaeology; Bayesian regression; palaeoclimatology; radiocarbon dating; radiocarbon-dated event count (REC) model

Introduction

Radiocarbon-dated event-count (REC) sequences are often used as proxies for important past human and environmental processes. Archaeologists, for example, use them as proxies for past changes in human demography, while palaeoclimatologists use them to understand phenomena such as past changes in sea level or natural fire regimes (e.g. Thorndyraft and Benito, 2006; Turney and Brown, 2007; Riede, 2009; Shennan, 2009; Bleicher, 2013; d'Alpoim Guedes *et al.*, 2016; Edinborough *et al.*, 2017). The underlying logic is based on a correlation between organic carbon deposition and a given past process of interest. In archaeology, the target process is often population-level change (e.g. Gamble *et al.*, 2005; Lepofsky *et al.*, 2005; Turney and Brown, 2007; Collard *et al.*, 2010; Schulting, 2010; Steele, 2010; Faulkner, 2011; Armit *et al.*, 2013; Shennan, 2013; Prentiss *et al.*, 2014; McLaughlin *et al.*, 2018; Colledge *et al.*, 2019; Hannah and McLaughlin, 2019; Leipe *et al.*, 2019). Spatio-temporal variation in human population levels is thought to be related to variation in organic carbon deposition because certain human activities lead to concentrations of organic carbon in sediment and those activities occur more often when and where there are more people present. Activities such as agriculture, construction, plant and animal processing, fire-use, and mortuary rituals all create localized concentrations of organic carbon—carbon deposition events. The frequency with which these events occur and the amount of carbon deposited are tied to population size (Rick, 1987). Thus, the number of radiocarbon samples in a given area dated to a given time can be expected to correlate with the number of

people who occupied that area at the relevant time. The logic is similar for palaeoclimatology (e.g. Pierce *et al.*, 2004; Thorndyraft and Benito, 2006; Mooney *et al.*, 2011; Bleicher, 2013). More natural fire events in a given place and time, for instance, lead to more charred carbon in the corresponding sediments. In both disciplines, the proxies are often compared to one or more potential covariates in an effort to explain through-time variation in the target process (e.g. Gamble *et al.*, 2005; Lepofsky *et al.*, 2005; Thorndyraft and Benito, 2006; Bleicher, 2013).

Despite the simple logic, however, radiocarbon dating uncertainty complicates the construction and use of these proxies. Radiocarbon dates contain uncertainty both from the measurement of carbon isotopes and from the calibration process that accounts for through-time changes in the ratio of those isotopes in the environment (Taylor *et al.*, 2014). These sources of uncertainty combine to produce calibrated radiocarbon dates with substantial, highly irregular temporal distributions (e.g. Fig. 1). It is common for a single sample to have a multimodal temporal density that spans centuries of calendar time. As a result, count-based sequences of radiocarbon-dated events are chronologically uncertain, which means that multiple potential sequences are always possible for any given sample of radiocarbon dates.

The primary way this uncertainty has entered into analyses of radiocarbon date proxies involves summing the levels of radiocarbon date densities for each slice of time in a given interval. The resulting proxy is often called a 'summed probability density function' (SPDF), and it has been a popular tool since its introduction to palaeoenvironmental science (Geyh, 1969) and archaeology (Dye and Komori, 1992). The proxy does have important limitations, however, and has been the subject of serious scrutiny and much criticism (e.g. Bamforth and Grund, 2012;

*Correspondence: W. Christopher Carleton, as above.
E-mail: wcarleton@ice.mpg.de

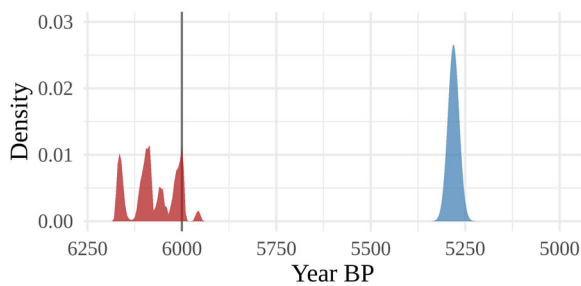


Figure 1. Example of an uncalibrated radiocarbon date density with its corresponding calibrated density. The vertical line represents the true underlying calendar date; the blue density represents the radiocarbon date density in radiocarbon years that would be returned from a dating laboratory; and the red density represents the calibrated radiocarbon date density with its domain in calendar years. [Color figure can be viewed at wileyonlinelibrary.com]

Williams, 2012; Bleicher, 2013; Kerr and McCormick, 2014; Manning and Timpson, 2014; Brown, 2015; Crema *et al.*, 2017).

One of the key problems is that individual radiocarbon date densities are an expression of uncertainty as just explained – like any measurement with error—which means sums of such densities are too. The level of a radiocarbon-date density for a given interval (say a year or a decade) is proportional to the probability that the relevant event occurred in the given interval (Bronk Ramsey, 2009). The sum of the levels of two such densities in the same interval, then, reflects both the number of events (because there are two events in question) *and* a function of the probability that each occurred in the relevant interval. When used as a proxy for the former, however, the sum is conflating process variation (the number of events) with measurement uncertainty (through-time fluctuations in the relevant probabilities). To further complicate matters, calibrated radiocarbon dates also reflect uncertainty in the calibration curve, which means the final sum of calibrated densities does as well (Brown, 2015). This is not necessarily a problem if, following Bronk Ramsey (2017), the proxy is only interpreted as the temporal distribution of the total chronological information in a radiocarbon dataset. The desired proxy in most cases, however, is the number of events that occurred in each interval of a sequence (e.g. Hoffmann *et al.*, 2008; Broughton and Weitzel, 2018; McLaughlin *et al.*, 2018) and the SPDF cannot isolate that number from chronological uncertainty about the timing of the individual events.

This conflation gives rise to analytical problems. It leads to proxy features that are unrelated to event-count variation, and these features are easy to misinterpret as being indicative of significant changes in the target process (Bamforth and Grund, 2012; Brown, 2015). It also means that information about uncertainty is being treated like information about process variation. In statistical terms, this information misplacement is a kind of model mis-specification that is likely to create inferential problems just as with other better-studied kinds of mis-specification that have been known about for a long time (e.g. Ramsey, 1969). Thus, for both qualitative and quantitative studies the information misplacement could lead to the identification of spurious relationships with covariates and incorrect confidence estimates. This is a significant problem that severely limits our ability to understand important long-term human and environmental dynamics with archaeological and palaeoenvironmental data.

There have been several recent attempts to improve the handling of uncertainty with respect to radiocarbon-date proxies. These improvements can be roughly divided into two groups. The first focuses primarily on null hypothesis testing. For the approaches in this group, the SPDF is still used, but observed data are compared either to SPDFs of dates simulated from theoretical growth models (Shennan *et al.*, 2013; Manning and

Timpson, 2014; Crema and Kobayashi, 2020) or to a null distribution derived from the observed data with a permutation procedure (Crema *et al.*, 2017, 2016). The former attempts to separate the target process from well-known spurious features introduced by the calibration process (Brown, 2015), while the latter attempts to account for spurious patterns produced by sampling variability (variation in spatial sampling intensity, specifically). In contrast, the second group focuses on improving the way a set of date densities are summarized. This includes methods such as sample bootstrapping (McLaughlin, 2019), Bayesian Gaussian mixture models (unpublished function in BChron, an R package for Bayesian radiocarbon date calibration; <https://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 very smooth estimates. The KDE approaches also produce uncertainty envelopes that can help, at least visually, to discern important variation from spurious fluctuations caused by the calibration curve (Bronk Ramsey, 2017).

While these approaches come with advantages, they are hindered by important disadvantages when the ultimate aim is to use radiocarbon-dated events as a proxy for some underlying process such as population-level change. The null hypothesis methods still rely directly on SPDFs and so conflate process variation with chronological uncertainty [though, see Crema and Kobayashi (2020) for a comparison between an SPDF and an alternative demographic proxy instead]. The new average/composite density estimation methods have fewer spurious features than SPDFs because they reduce calibration artefacts, but they too are summaries that mix up process variation and chronological uncertainty. An important caveat with respect to the KDEs, however, is that they are based on a set of sub-models that individually do not conflate process variation with chronological uncertainty. The sub-models are also KDEs and they are combined to produce an average (Bronk Ramsey, 2017) or composite (Brown, 2017) model. Individually, these sub-models do not conflate process variation with chronological uncertainty because each one represents the temporal density of a single randomly sampled set of event times drawn from the underlying radiocarbon date densities. When combined, however, the level of the resulting composite/average KDE at a given time is a weighted sum that expresses both the temporal density of events and uncertainty about the timing of those events. Thus, the information misplacement described earlier persists with both the extended SPDF approaches and the primary KDE models. Consequently, they may not be suitable for parameter estimation, model comparison or variable selection when the scientific objective is understanding through-time changes in the number of events while accounting for chronological uncertainty.

Here I describe a study in which I explored an alternative Bayesian approach inspired by multilevel modelling. In a typical multilevel framework, parameter estimates for a given variable are thought of as samples from a common super-population distribution (Fig. 2) (Gelman and Hill, 2007; Gelman *et al.*, 2013). For example, one might conduct several regression analyses, each involving a different sample of observations. The key parameter of interest might be the relationship between a given covariate and a response variable, such as calories consumed and weight gained, respectively. The relationship is characterized by a regression coefficient, β_j , where j refers to a specific data sample (e.g. sample of individuals whose caloric consumption and weight have been measured). Each β_j could plausibly be thought to come from some super-population of β 's because we expect variability between data samples to influence the estimate of any individual β_j . The β_j 's, therefore, can be modelled with a

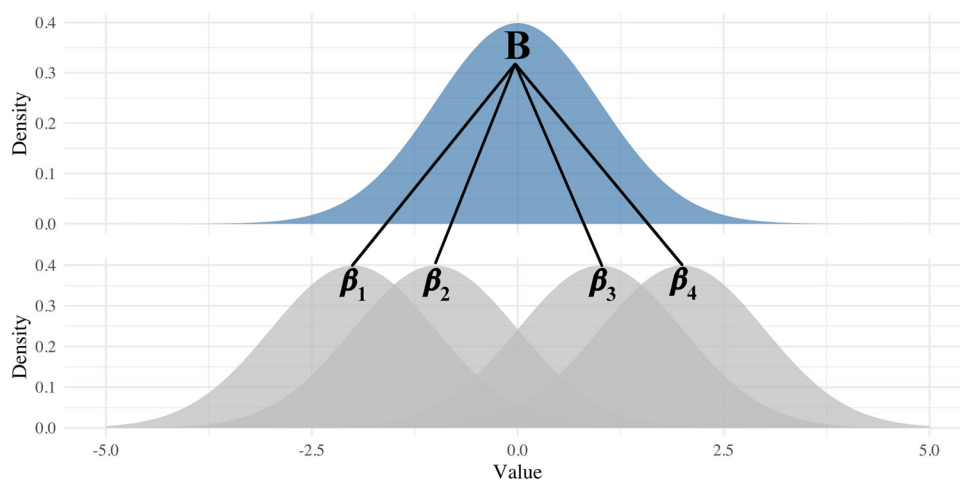


Figure 2. Schematic of a multilevel model. The blue density in the top panel represents a superpopulation parameter and the grey densities in the lower panel represent samples of that superparameter. [Color figure can be viewed at wileyonlinelibrary.com]

probability distribution that describes their uncertainty with respect to variability among samples. That prior has its own mean, B , and sampling uncertainty (standard deviation), σ_B . With several samples (a number of j 's) Bayesian methods can be used to estimate both the individual β_j 's and the super-population parameters, B and σ_B .

A similar approach could be used to account for chronological uncertainty in a RECE sequence. Instead of exploring sampling variability among distinct sample clusters, however, chronological uncertainty could be accounted for by imagining that every possible event-count sequence is like a sample from a super-population. Take, for instance, two radiocarbon-dated events with overlapping temporal densities. With only two densities, the set of possible event-count sequences

includes realizations wherein the events do not co-occur and others wherein they do (Fig. 3). Sequences of the first kind contain two instances where the count is equal to 1, corresponding to the dates in which each event occurs, and zeros elsewhere (Fig. 3C). In other sequences, the corresponding event-count is equal to 2 at the time of co-occurrence and 0 elsewhere (Fig. 3D). Importantly, there are also enormous numbers of possible permutations because the specific date for the co-occurrence (or, conversely, separate occurrences) could be any of the potential dates within the interval spanned by both densities. Without a compelling reason to assume that one particular sequence is the 'true' one, all the possibilities need to be explored. Thus, we can imagine that all the potential sequences make up a super-population of sequences

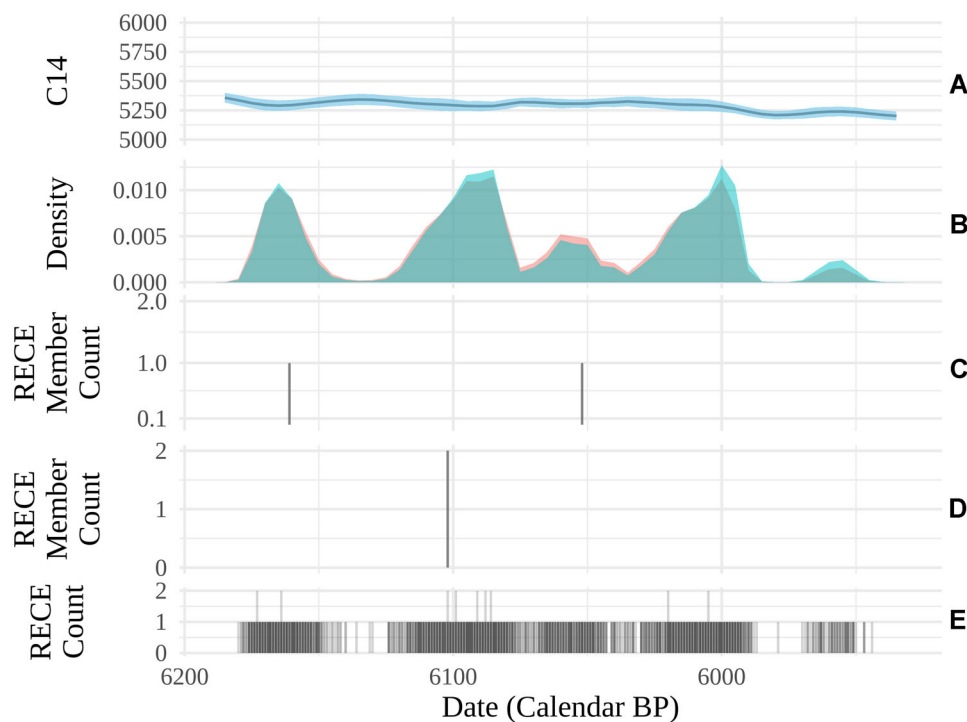


Figure 3. Example of the effect of chronological uncertainty on event-counts. Panel A shows the IntCal13 calibration curve; B shows two overlapping calibrated radiocarbon date densities – the true calendar dates are not shown because of the scale (the dates are 6000 and 5999 BP); C shows an example probable event-count sequence in which the events do not occur in the same year; D shows an example sequence in which they do occur in the same year; and E shows a randomly drawn ensemble of probable event sequences (RECE) plotted with transparency so that higher density regions can be visually distinguished (darker areas). [Color figure can be viewed at wileyonlinelibrary.com]

reflecting the chronological uncertainty in the timing of the individual events. This approach deviates from true multilevel modelling because the units of analysis (potential REC sequences) are not ontologically distinct like separate nested clusters of observations (e.g. students in different schools or patients in different hospitals, canonical examples of multi-level data). Nevertheless, viewed in this way, the super-population can reflect the impact of chronological uncertainty on various parameters of interest. If, for example, we were interested in the relationship between the true event-count sequence and one or more potential explanatory covariates, the super-population could be sampled to estimate that parameter and account for the chronological uncertainty in the timing of the relevant events. The model would summarize the results of a set of what-if scenarios in which we ask what the model parameters would be if a given event-count sequence represented the true underlying process and then permuted all the sequences—though, in practice, we would randomly sample a set of probable sequences (creating a Radiocarbon-dated Event Count Ensemble, or RECE) rather than permute all possibilities. In this way, the approach extends the KDE (Bronk Ramsey, 2017) and CKDE (Brown, 2017) approaches by introducing a Bayesian regression framework instead of producing a composite/average density estimate. This new approach—REC modelling—could hypothetically be used for many different types of models and parameters of interest.

To explore the potential utility of REC models, I ran a series of simulated regression experiments in two stages. The aim was to see whether a known relationship between a given simulated covariate and REC sequence could be reconstructed with a multilevel Bayesian regression model designed to account for chronological uncertainty. The first stage involved a simple monotonic function representing persistent through-time changes in event frequency, and the second stage involved a real palaeoclimate record representing highly variable through-time changes in event frequency. In all regression experiments, the true underlying process—time in the case of the monotonic function in the first stage and the palaeoclimate record itself in the second—was also used as the sole covariate. I reasoned that if

the proposed approach could be used in a general applied setting, it should be able to produce good estimates of the known regression coefficient. A ‘good’ estimate in this context would be one that was at least close to the true value, indicating the correct direction and magnitude of the known effect. If the estimates were biased, however, the method would still be considered potentially useful if the bias was in a consistent direction (i.e. always too high or too low, irrespective of the structure of the underlying process). Importantly, the method would only be considered useful if it produced good estimates in the analyses of both the monotonic function *and* the real palaeoclimate record.

Materials and methods

Simulated data

As mentioned, two processes were used to simulate data (Figs 4–7). The first was a monotonic function of time. This process could represent, for example, a simple through-time increase/decrease in human population levels or forest fire frequency. It was formulated as follows:

$$y_t = e^{\beta t} \mid t \in \{t_1, t_2, \dots, t_n\} \quad (1)$$

where β represents a regression coefficient (exponential rate) and t refers to a given date from t_1 to t_n . The start date (t_1) was the beginning of a millennium between 12 000 and 2000 a BP chosen at random, and the end date was 1000 years later—the values used were $t_1 = 6000$ a BP and $t_n = 5001$ a BP. The regression coefficient, β , was set to 0.004 for one set of analyses and then to -0.004 for another set.

For each analysis in this stage of the study, the process in Equation (1) was used to produce a sample of 1000 calendar dates. These dates were then turned into an event-count sequence by counting the number of dates that fell into a given year from 6000 to 5001 a BP. Next, the sample of calendar dates was converted into corresponding radiocarbon dates (on the ^{14}C timescale) with a function in the ‘Clam’ R package (Blaauw, 2020) and then calibrated back onto the calendar

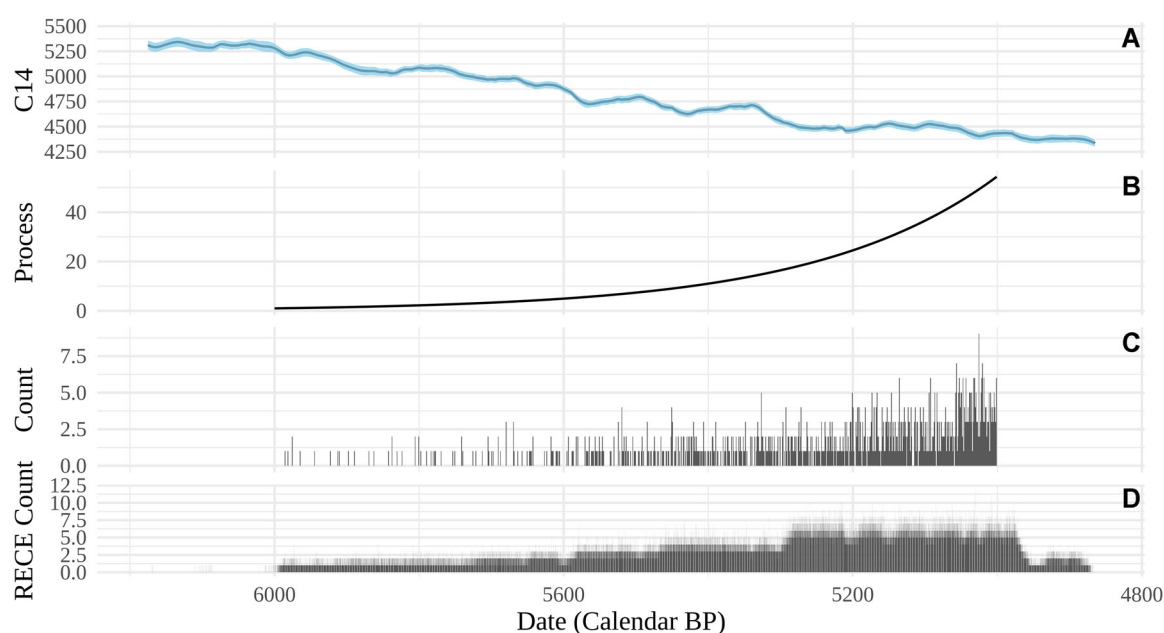


Figure 4. Plots showing simulated data for the monotonic *increasing* process. Panel **A** displays the IntCal13 calibration curve; **B** shows the true process; **C** shows the sample of 1000 exact dates from the process; and **D** shows a radiocarbon-dated event-count ensemble (RECE) of the process. [Color figure can be viewed at wileyonlinelibrary.com]

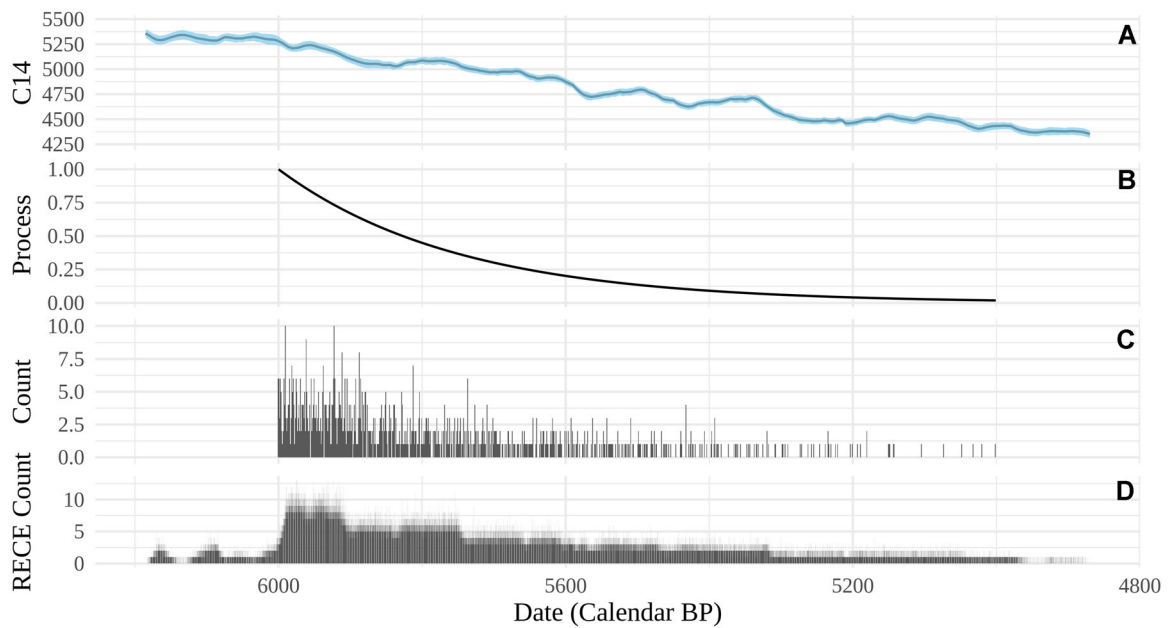


Figure 5. Plots showing simulated data for the monotonic *decreasing* process. Panel **A** displays the IntCal13 calibration curve; **B** shows the true process; **C** shows the sample of 1000 exact dates from the process; and **D** shows a radiocarbon-dated event-count ensemble (RECE) of the process. [Color figure can be viewed at wileyonlinelibrary.com]

scale with the same package. Lastly, the set of calibrated radiocarbon-date densities was sampled at random to produce an ensemble of potential sequences. The dates in the domain of the densities were sampled in accordance with the level of the given density, which meant that the sampled ensemble reflected the probability that each possible event-count sequence would be observed given the set of radiocarbon dates (i.e. more probable sequences were sampled more often).

The second simulated process was based on a real palaeoclimate record. The record chosen was a high-resolution time-series of ^{14}C measurements from a speleothem extracted from Yok Balum Cave in southern Belize (Kennett *et al.*, 2012). The series was chosen because it has a large

number of observations. The specific period of real time actually covered by the series was not relevant, however, and neither was its temporal resolution. Rather, it simply represents a natural process with the kind of variability observed in real climate data, which made it a useful basis for this simulation study. The observations in the speleothem record were used as the covariate, x , in the following regression expression:

$$y_n = e^{\beta_0 + \beta x_n} \quad (2)$$

where β and β_0 were again regression coefficients. This function (Equation 2) was then used to weight the probability that a given integer date in the range [6000,5001] a BP—chosen to be consistent with the previous stage of the

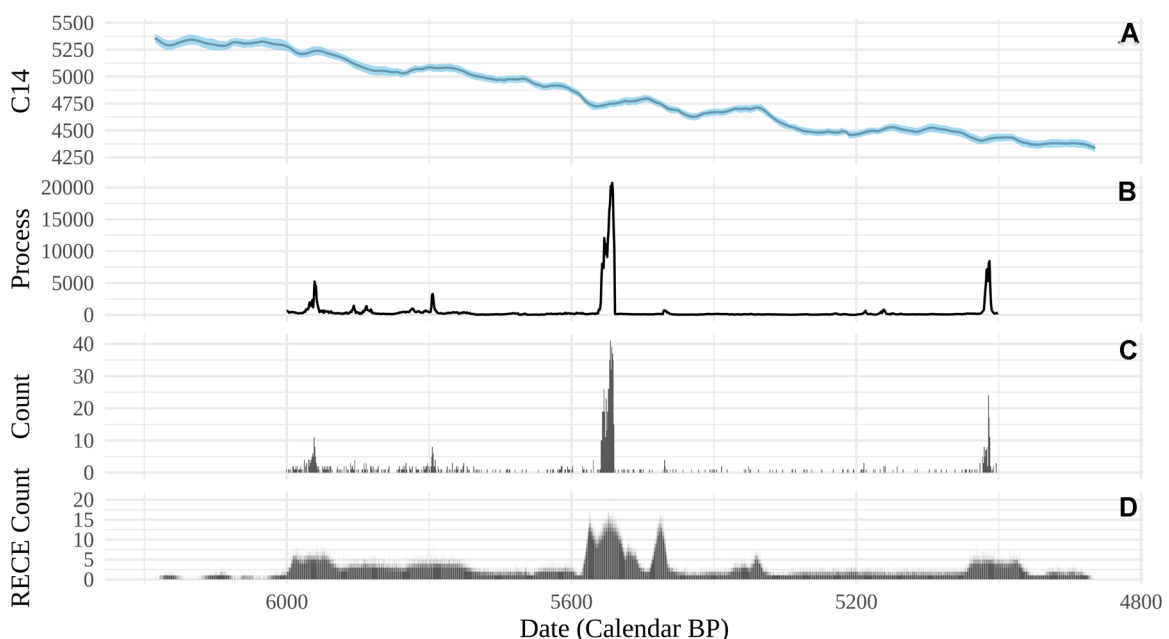


Figure 6. Plots showing simulated data based on a process driven by a *positive* relationship with a real palaeoclimate record. Panel **A** displays the IntCal13 calibration curve; **B** shows the true process; **C** shows the sample of 1000 exact dates from the process; and **D** shows a radiocarbon-dated event-count ensemble (RECE) of the process. [Color figure can be viewed at wileyonlinelibrary.com]

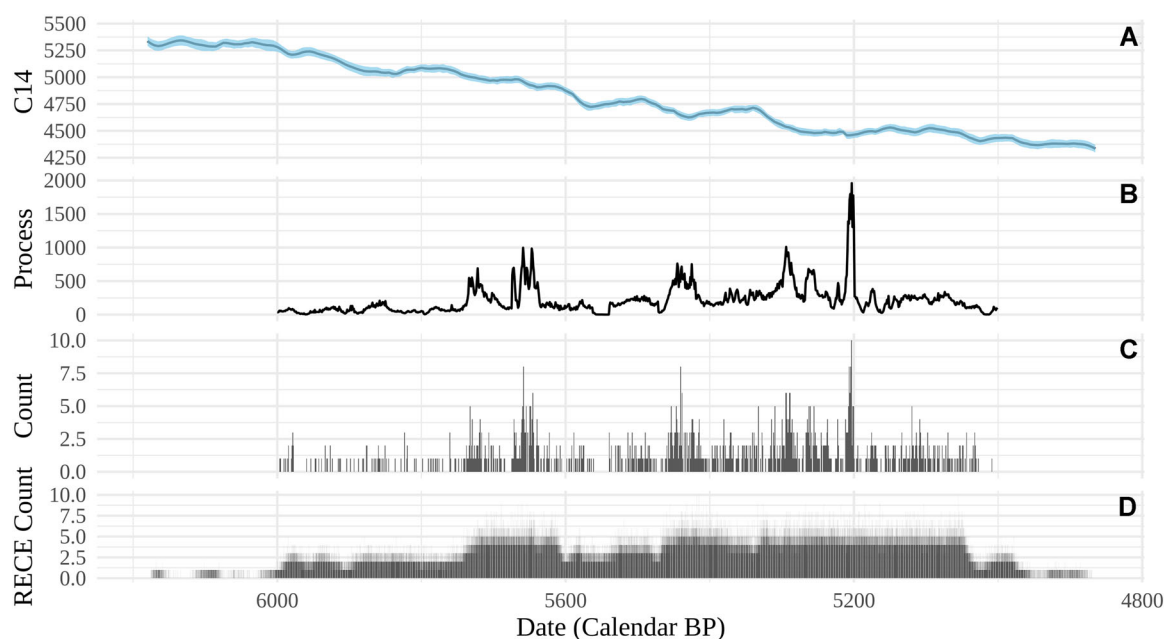


Figure 7. Plots showing simulated data based on a process driven by a *negative* relationship with a real palaeoclimate record. Panel **A** displays the IntCal13 calibration curve; **B** shows the true process; **C** shows the sample of 1000 exact dates from the process; and **D** shows a radiocarbon-dated event-count ensemble (RECE) of the process. [Color figure can be viewed at wileyonlinelibrary.com]

study—would be randomly selected. When the speleothem record was high, the corresponding date would be selected with a higher probability, and vice versa. An R core function ‘sample’ (R Core Team, 2019) was used to perform the weighted sampling and produce 1000 dates. Next, the sampled calendar-scale dates were converted into radiocarbon dates and calibrated following the same procedure just described for the monotonic function. As before, this procedure was followed twice to explore both a positive and a negative regression coefficient, which was set to ± 1 . In this case, to compensate for large magnitude negative numbers in the climate record, an intercept term was added to the model, β_0 , and set to 5 for all analyses. This intercept term ensured that the mean level of the count process was high enough that the sampled sequences would have sufficient variation in them – without it, the mean level of the process was so close to zero that most of the sampled sequences were largely just zeros.

Regression experiments

Three regression analyses were performed in each stage of the study. In the first analysis of each stage, the response variable was the event-count sequence with no chronological error. This analysis served to verify that without chronological uncertainty the known relationship between the count sequence and its corresponding process—the monotonic function or the real palaeoclimate data—could be accurately estimated. In the second analysis, a single sequence from a given ensemble was used as the response variable. Then, in the last analysis, a randomly selected sample of 50 sequences from the relevant ensemble was analysed within a multilevel framework wherein each sequence served as the response variable in a single regression (i.e. a full REC model).

In all analyses the relevant true process served as the sole covariate. Thus, time was used as the sole covariate for the monotonic process in stage one, and the palaeoclimate record was used as the sole covariate in stage two. The target parameter in each case was the relevant known regression coefficient.

Radiocarbon-dated event-count (REC) model

The core model used in all analyses was based on the negative binomial distribution (Hilbe, 2011). This distribution was chosen for two reasons. First, its domain is the positive integers, which is appropriate for event-count sequences. Second, it is often used in cases where the relevant response variable may have a variance/dispersion that is not identical for all values of a given covariate (e.g. Zhu, 2011). This property makes it useful for capturing autocorrelation and non-stationarity—and, in the present case, it was useful for modelling a particular property of chronological uncertainty.

Chronological uncertainty imposes a specific structure on the set of all possible event-count sequences. That structure is a gradual increase followed by a gradual decrease, a kind of spreading along the time axis in both directions (Bronk Ramsey, 2017). As Figs. 4 and 5 show, even if the true process increases or decreases monotonically with time, the set of possible event-count sequences will always gradually increase and then gradually decrease. The backward and forward smearing of the true process is caused by the natural structure of chronological uncertainty. When we say we are uncertain about the timing of an event, we really mean that we have some idea about when the event occurred and that the probability of the event having occurred in any particular time declines with distance from some central, most-probable interval. When compiling an event-count sequence, some events will sit at the ends of the relevant temporal interval. Consequently, portions of the densities describing the chronological uncertainty of those leading and trailing events will extend beyond most of the densities in the event sample. This tapering creates gradual tails in the sequences. Without the true process to indicate otherwise, we must assume that those tails, however small, indicate possible temporal locations for some events. Consequently, the number of events that probably occur in those tail regions declines gradually away from the mass of most event densities. Even single potential sequences randomly selected from an ensemble will tend to display some level of smearing—‘chronological spread’ (Bronk Ramsey, 2017)—because of the nature of chronological uncertainty. Importantly, the smearing does not

reflect the true underlying process and, so, can potentially confound statistical analyses that fail to account for it.

The negative binomial distribution may be able to account for the spread. The distribution has two parameters. One, often denoted r , refers to some number of events (either failures or successes) in a set of Bernoulli trials (coin-flips); and the other, p , refers to the probability of the event (success or failure). The domain of the distribution then refers to the corresponding number of complementary events that would be expected for a given combination of r and p . For example, if r refers to successes, then the domain is often the expected number of failures for a given number of trials. Importantly, the mean of the distribution can be written as follows:

$$\mu = r \frac{1-p}{p} \quad (3)$$

In this equation, p acts like a weight applied to r . Thus, r could refer to a regression function intended to capture the impact of covariates on the mean, and p would then be useful for capturing through-time changes in the mean that may be unrelated to the covariates. In the case of REC sequences, p could represent the expected impact of chronological uncertainty on a given sequence. It would adjust the mean level of the process distribution up or down in accordance with gradual increase/decrease in a given sequence caused by the structure of chronological uncertainty (i.e. the gradual forward and backward tapering of the event sequence as described earlier).

This view of the negative binomial distribution's parameters led to the core regression model used for all analyses. The model was structured as follows:

$$y_n \sim NB(r, p_n) \quad (4)$$

$$r = e^{\beta_0 + \beta x_n} \quad (5)$$

The outcome (response) variable in Equation (4) is denoted y_n , which refers to the level of a given event-count process at time n . That variable follows a negative binomial distribution, $NB(\cdot)$ and has a parameter r that must be positive and p , which must be between zero and one. In Equation (4), r is defined by a log-link function containing a regression with an intercept denoted β_0 , a regression coefficient β and a covariate x_n . In practice, of course, the model could be adapted to include any number of potential covariates by turning β into a vector and x_n into a matrix where the columns are each a different covariate. The model is Bayesian and, so, has the following priors defined for the random parameters:

$$p_n \sim U(1e^{-10}, 1) \quad (6)$$

$$\beta_0 \sim N(0, 100) \quad (7)$$

$$\beta \sim N(0, 100) \quad (8)$$

To account for chronological uncertainty in event times, Equation (4) was extended. Instead of only estimating one set of parameters for a single regression involving one event-count sequence (y_n), multiple regressions were set up and their parameters estimated simultaneously for k potential event-count sequences, denoted $y_{n,j}$. Hyperpriors were added for the super-population regression coefficient and the intercept denoted B and B_0 , respectively. The lower-level individual regression parameters were denoted β_j and β_{0j} . The whole model (i.e. full REC model) was structured as follows:

$$y_{n,j} \sim NB(r_j, p_{n,j}), \quad (9)$$

$$r_j = e^{\beta_{0j} + \beta_j x_n} \quad (10)$$

where x_n still referred to a single covariate—the true process responsible for the event-count sequence. The model also had the following priors:

$$p_{n,j} \sim U(1e^{-10}, 1) \quad (11)$$

$$\beta_{0j} \sim N(B_0, \sigma_{B_0}) \quad (12)$$

$$\beta_j \sim N(B, \sigma_B) \quad (13)$$

$$B_0 \sim N(0, 100) \quad (14)$$

$$B \sim N(0, 100) \quad (15)$$

$$\sigma_{B_0} \sim U(1e^{-10}, 10) \quad (16)$$

$$\sigma_B \sim U(1e^{-10}, 10) \quad (17)$$

All the analyses were conducted in R (R Core Team, 2019). A Bayesian analysis package, 'Nimble' (NIMBLE Development Team, 2018), was used to produce the Bayesian models and estimate their parameters with Markov chain Monte Carlo (MCMC) simulations. Each simulation involved a minimum of 200 000 iterations. The MCMC posterior chains were inspected for diagnostic indications of convergence, and standard Geweke statistics (Geweke, 1992) were used to test the chains for significant deviations from stationarity after discarding the first 2000 iterations as burn-in (see Supporting Information for Geweke test results). The packages 'ggplot2' (Wickham, 2016), and 'ggpubr' (Kassambara, 2019) were used for producing figures. All the R scripts used in this analysis are available at www.github.com/wccarleton/recm.

It should be noted that the sequences analysed were sub-sampled to speed up parameter estimation. Each model could involve thousands of parameters. This is because the average ensemble member event sequence originally contained roughly 1300–1500 observations and each observation had a corresponding p parameter to be estimated. With a sample of only 50 ensemble members in a given model, then, the number of p parameters could be as high as 75 000 per analysis. Such an enormous number of model parameters would require millions of MCMC iterations to obtain unbiased posterior samples. So, to shorten computation time and reduce memory requirements, the response/covariate pairs were sub-sampled such that only every 10th observation was included in a given lower level regression—a few thousand parameters to estimate instead of tens of thousands. For similar reasons, only 50 sequences were drawn from the parent population of possible sequences (but see the Supporting Information for replications involving more sequences). While this may not be recommended in general, the simulated patterns were very clear by design, which meant that sub-sampling had no obvious impact. In practice, this would reduce the likelihood that a noisy pattern and/or weak signal could be detected.

Results

Stage 1: monotonic process

The results from the first stage of the study were promising. To reiterate, this stage involved a monotonic function of time as

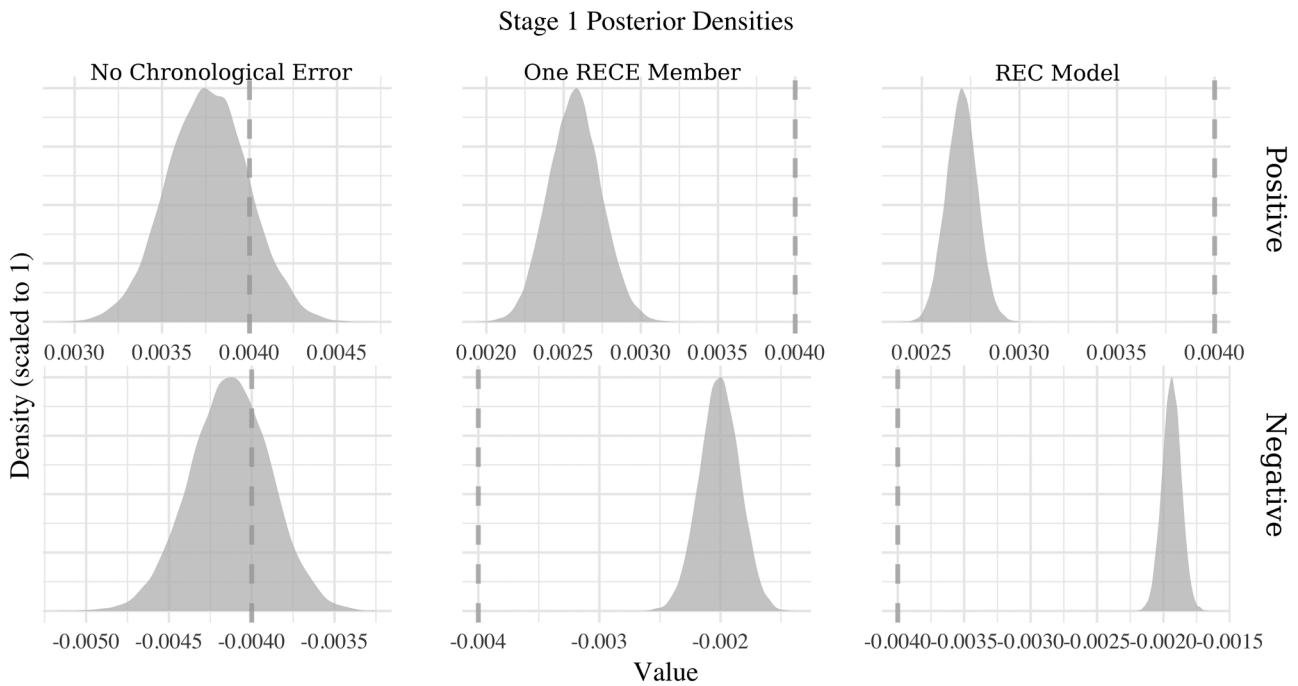


Figure 8. Density plots for posterior estimates of the target regression coefficient in each of the three analyses of the first stage (i.e. the analyses in which a monotonic function was used as the true process). The far left column shows posterior densities for the β parameter in the models involving the chronological-uncertainty-free date sample. The middle column shows the same posterior estimates for the models involving a single randomly selected RECE member. The far right column contains posterior density plots for the top-level, superpopulation B parameters estimated using an REC model.

the underlying process and the target parameter was a regression coefficient set at ± 0.004 . The first analysis—the verification with a chronological-uncertainty-free event sample—produced a mean posterior estimate of 0.0041 for the target parameter when the true effect was positive and -0.0039 when it was negative. This result demonstrated that the core negative-binomial regression model was able to correctly estimate the target parameter with a reasonable degree of uncertainty (Fig. 8). As Fig. 8 shows, the model also produced a reasonable posterior predictive interval that included all the observed data in the 95% interval—the interval was estimated with point-wise quantiles of posterior samples from the MCMC.

The next regression analysis involving a single REC sequence as the response variable also produced promising results. The posterior mean estimate of the target parameter was 0.0025 when the true effect was positive and -0.0021 when it was negative (Fig. 8). These estimates indicate the correct direction of the known effect, but both are biased towards zero. Importantly, the posterior predictive estimates for the observed data also included all the data points at the 95% credible interval despite the temporal smearing caused by chronological uncertainty (Fig. 9).

The full REC model results were similar. Recall that in this analysis 50 randomly sampled event-count sequences were simultaneously assessed and that the individual regression

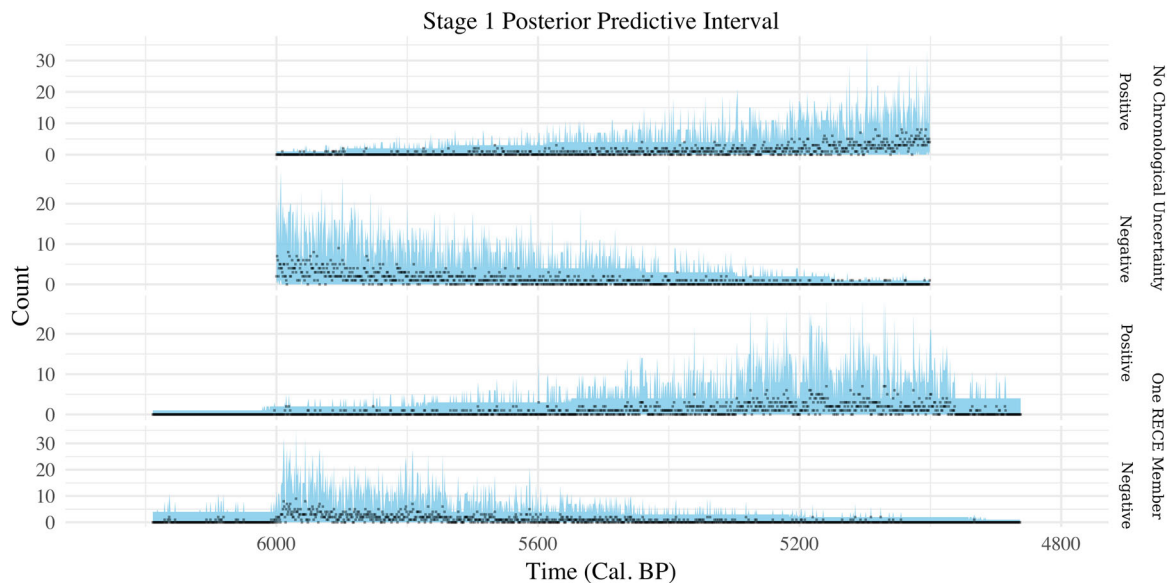


Figure 9. Posterior predictive intervals for the models of the monotonic function. The black dots represent the observed count data and the blue region represents the 95% predictive interval. [Color figure can be viewed at wileyonlinelibrary.com]

coefficients, β_j 's, from those analyses were considered samples of a super-population parameter, B . Thus, the estimates of interest in this case were primarily the super-population parameter, B , and its sampling uncertainty, σ_B . The mean of the posterior distribution for B was approximately 0.0027 when the true relationship was positive and -0.0018 when it was negative. The mean of the posterior distribution for B 's sampling uncertainty (standard deviation), σ_B , was 0.000076 when the relationship was positive and 0.000057 when it was negative.

Stage 2: real palaeoclimate process

The results from the second stage of the study were also encouraging. In this stage, a real palaeoclimate record was used to create the process that drove event-counts. The target parameter was a regression coefficient set at ± 1 . The first analysis—the verification—produced a posterior mean estimate of 1.009 for the target parameter when the true effect was positive and -0.989 when it was negative. As with the corresponding analysis in the first stage of the study, the model produced a reasonable posterior predictive interval that included all the observed data at the 95% level.

The next regression analysis involving a single REC sequence also produced promising results. The posterior mean estimate of the target parameter was 0.558 when the true effect was positive and -0.459 when it was negative (Fig. 10). Like the stage 1 results, these estimates have the correct signs but are biased towards zero. The posterior predictive estimates for the observed data included all the data points at the 95% credible interval despite the temporal smearing (Fig. 11).

The full REC model produced similar results. The posterior mean for β —the super-population regression coefficient—was approximately 0.5 when the true relationship was positive and 0.5 when it was negative. The mean of the posterior distribution for the sampling uncertainty, σ_B , of the super-population regression

parameter was 0.02 when the relationship was positive and 0.037 when it was negative.

Discussion

The simulation results indicated that REC models may have substantial utility. In all the experiments, the models produced posterior parameter estimates that were consistent with the sign of true values and biased towards zero. This means that the model always indicated the correct direction of the known effect and, although biased, was off in a consistent way. In practice, this means that REC models appear to be able to determine whether a real effect was positive or negative despite radiocarbon-date uncertainty and that in general we can expect the magnitude of effect estimates to be lower than their true values. While a biased estimate is suboptimal, the consistency of the bias suggests REC models are still useful for resolving important features of real response–covariate relationships even in the presence of significant chronological uncertainty.

The bias was probably caused by temporal spread. Temporal spread, as explained earlier, describes a distortion of the true event-count sequence caused by chronological uncertainty. Compared to the true event-count sequence, the radiocarbon-dated event-count ensemble (RECE)—set of probable event-count sequences—is distorted in two ways. One is that the count sequence is, on average, flattened with its peaks lower than those of the true sequence. This is because the probability that events co-occurred is almost always going to be lower than the probability that they occurred at different times. The other distortion involves overall temporal span. Compared to the true sequence, the RECE occupies a longer span of time and tapers at either end. It tapers backward and forward from the first and last observations in the true sequence, respectively. Thus, the RECE as a whole is flattened and spread out compared to the true event-count sequence.

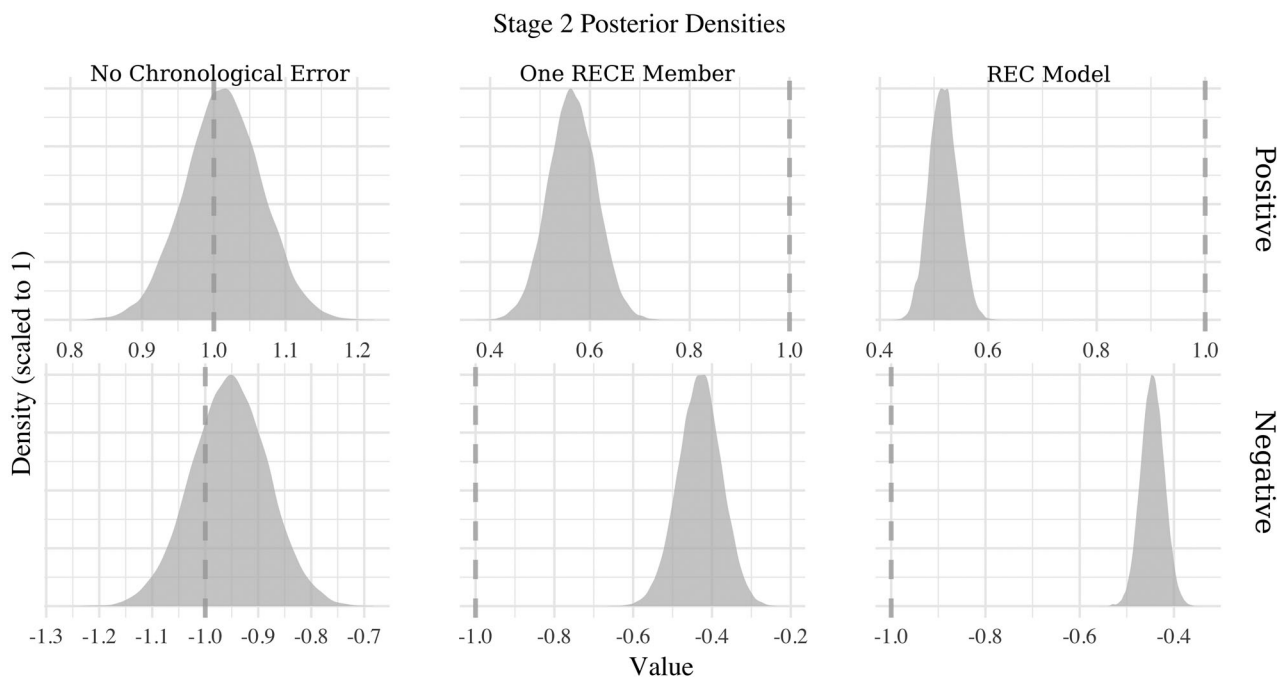


Figure 10. Density plots for posterior estimates of the target regression coefficient in each of the three analyses of the second stage (i.e. the analyses in which a real palaeoclimate record was used as the true process). The far left column shows posterior densities for the β parameter in the models involving the chronological-uncertainty-free date sample. The middle column shows the same posterior estimates for the models involving a single randomly selected RECE member. The far right column contains posterior density plots for the top-level, superpopulation β parameters estimated using an REC model.

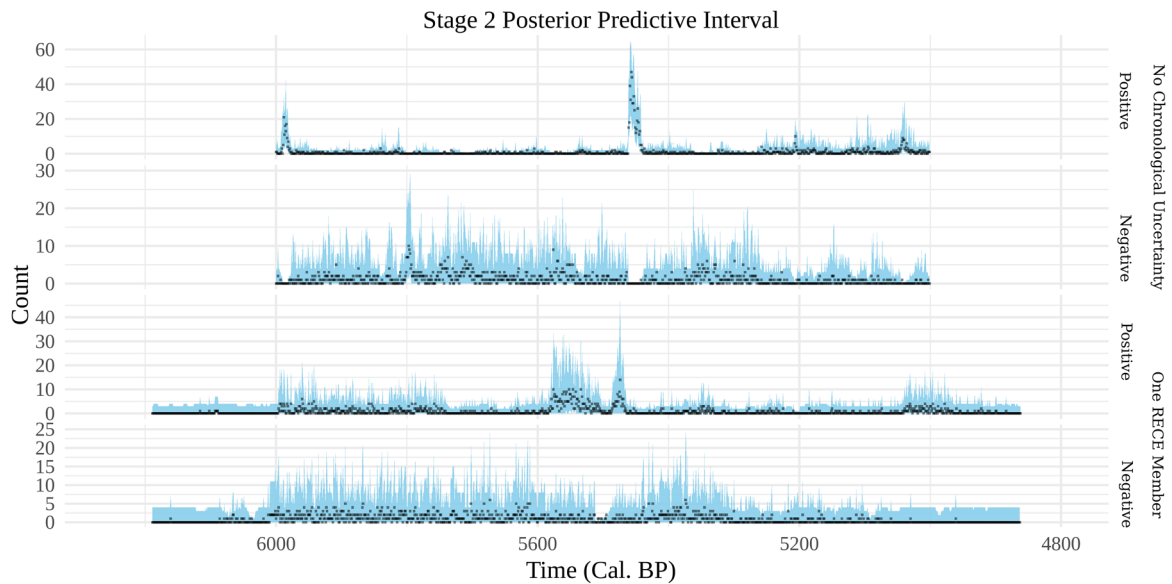


Figure 11. Posterior predictive intervals for the models based on the real palaeoclimate data. The black dots represent the observed count data and the blue region represents the 95% predictive interval. [Color figure can be viewed at wileyonlinelibrary.com]

As a corollary, a randomly drawn sequence from the RECE (representing one probable event-count sequence) will probably have lower peaks than the true sequence and the events will be spread over a longer interval.

The way these distortions lead to biased estimates can be seen most clearly in the case of a monotonic process (Fig. 12). The rate parameter for the monotonic function would naturally be lower for event-count sequences that have been squashed and stretched out in time. This is because the rate parameter is essentially a slope value with event-count as the numerator and time as the denominator. Dilation along the time dimension (i.e. spreading out of the events) increases the denominator, which means the ratio is lower for a given count sequence. Imagine the time between counts growing while the counts themselves remained constant. A line drawn through those counts would tilt, becoming more horizontal (i.e. have a

lower slope) as the time dilation increased. The decrease in the rate would be even greater if the numerator (the counts) shrunk as well, which as described earlier is what happens as a result of temporal spread. Thus, temporal spread may be expected to reduce the magnitude of regression coefficients with respect to time in many real-world cases. When time is the covariate, we should expect coefficient estimates to be closer to zero than their true values.

While harder to intuit, the bias is the same for regressions with respect to any covariate, x_t , that is itself a function of time whether it is monotonic or not. In such cases, the temporal spread of the event-count sequence still results in flattening and dilation of the sequence with respect to both the covariate and time, as long as the covariate observations are temporally fixed. The events, and corresponding counts of events per unit time, are spread out such that they become paired with a greater span

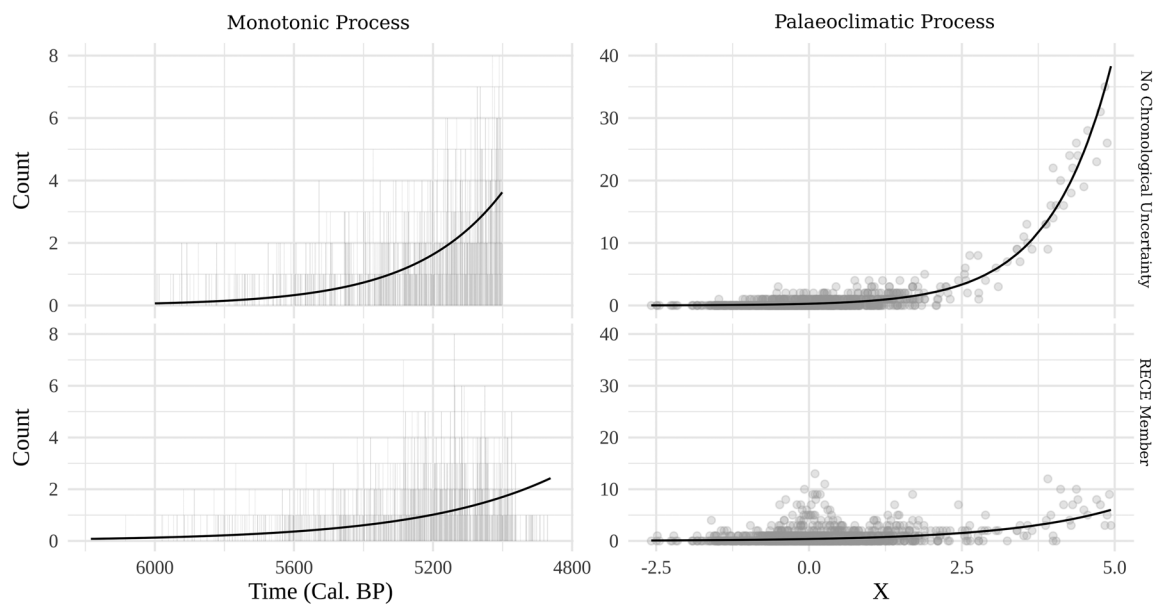


Figure 12. Plots demonstrating the effect of chronological spread on the rate of the monotonic function. The top two plots show the true function (black line) and corresponding chronological-error-free date samples from that function. The bottom two plots show the best approximation to the true function (black lines) given samples from an RECE. The RECE samples represent possible event-count sequences based on calibrated radiocarbon dates corresponding to the sample of chronological-error-free dates used in the top two plots.

of covariate observations compared with the original event-count sequence. If the covariate is the product of a stationary process (i.e. the covariate's process has statistical properties that remain fairly constant through time) then the event-count sequence is redistributed more evenly over roughly the same range of covariate values (as in Fig. 12). If, in contrast, the covariate process is not stationary, then the event-count sequence is spread out over a larger portion of the covariate's intrinsic variation as well. In both cases, the correlation between the count process and the covariate would be reduced relative to the true covariation. These distortions affect the 'slope' of the regression function with respect to the covariate just as they would with respect to time—i.e. the magnitude of the 'slope' is reduced while its sign remains the same.

Brown (2017) recently reported a similar bias with respect to regression involving SPDFs. In a simulation study analogous to the one described above, Brown used regression models to evaluate the standard SPDF and a CKDE model. Each summary proxy was created with simulated radiocarbon data based on a known process with a fixed rate parameter. Then, the proxies were each used as the response variable in a regression, one for the SPDF and one for the CKDE. As was found in the present study, the models Brown assessed produced parameter estimates that were closer to zero than the target value. Interestingly, the regression involving the SPDF produced a *less biased result* in the experiment than the one involving the CKDE, which is surprising given that KDE models ideally remove spurious radiocarbon calibration artefacts.

With this finding in mind, it seemed logical to compare the REC results to a simple regression involving an SPDF. So, following Brown (2017), I produced an SPDF and used R's built-in 'glm' function to estimate the parameters of a Gaussian generalized linear model (GLM) involving that SPDF as the response variable and time as the sole covariate. The SPDF was based on the simulated data from the monotonic (increasing) process used in the present study, which meant that the target rate parameter had a value of 0.004. The SPDF-GLM produced an estimate that was biased further towards zero than the REC model estimate of the same parameter (Fig. 13).

The advantage of the REC model may come partly from sampling possible event-count sequences in proportion to their relative probabilities. This probability-weighted sampling likely reduced the impact of temporal spread on the REC model compared to the SPDF-GLM because the latter effectively gives equal weight to every point on the SPDF including the 'tails' at either end of the proxy. Put another way, the SPDF-GLM cannot account for the difference between the probability that an event occurred in a given time and the probable number of events that occurred—it was misplacing information by treating chronological uncertainty as if it was direct information about event-counts. It should also be noted,

though, that a Gaussian model is probably not appropriate when applied to an SPDF (or KDE) because those are bounded on the low end by zero—they must be positive by definition and the Gaussian distribution does not reflect that. So, some of the additional bias observed in the GLM-SPDF estimate may have been partly caused by that distributional mis-specification as well.

The temporal spread bias must be considered when interpreting REC model parameters. It means that, in general, smaller effects will be harder to see, holding other factors constant (e.g. overall degree of chronological uncertainty in the events and the relevant portion of the calibration curve). Small effects that are already close to zero may become indistinguishable from zero because temporal spread presses the estimates of those effects further toward zero. Consequently, the bias also means that effect sizes will be underestimated. So, for example, an analysis of the long-term impact of climate change on human population sizes would probably estimate an impact that is *too low*. It will be important, then, for inferences drawn from REC models to be tempered in light of the bias caused by temporal spread.

It will also be important to consider the impact of sampling insufficiency on REC models. While the issue of sampling has been discussed previously with respect to radiocarbon date proxies (e.g. Williams, 2012), REC models will probably be affected in a predictable way that bears stating explicitly. Count models, like the REC model presented here, can be affected by zero-inflation. Zero-inflation occurs when there is an excess of zeros in a count dataset. The excess can occur for many reasons, and it ultimately leads to biased parameter estimates and misleading inferences (Hilbe, 2014). For REC models, this problem implies that spatiotemporal sampling intensity needs to be considered carefully. In an extreme case, a sequence composed entirely of zeros because the relevant region or time period had never been investigated would appear to the REC model as representing a constant level (zero-level) of the target process. However, there would be an overabundance of zeros in this hypothetical dataset caused by insufficient sampling and not by a constant low level of the real target process. This cartoon example would obviously lead to a false impression. Less extreme examples, though, may be harder to detect in practice and REC model users need to be aware of the potential pitfall. Handled incautiously, poor sampling will probably produce misleading results and the sampling intensity required for a given study will be case-dependent. Possible solutions exist, such as zero-inflated Poisson models or mixture models (see Hilbe, 2014) but results may vary and the appropriateness of a given solution is again case-dependent.

A related sampling concern involves the notion of sample size and significance. Converting radiocarbon-date databases into count time-series transforms sample size estimates from

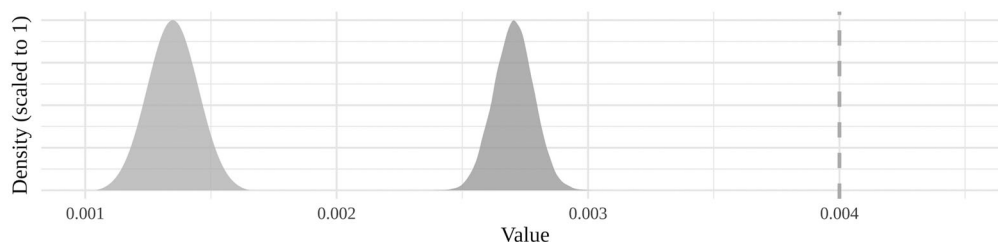


Figure 13. Parameter estimates from the SPDF-GLM and REC models compared to the true target value. The grey dashed line on the right indicates the target value. The centre dark grey density is the posterior density estimate for that target from an REC model. The light grey density on the left is the density of the SPDF-GLM estimate (constructed using the mean and standard error reported by R's 'glm' function) of the target parameter. As the plot makes clear, while both estimates are biased towards zero, the REC model estimate is much less biased than the SPDF-GLM estimate.

the number of dated events to the length of the series. For example, 100 dates spread out over 1000 years in an annual count series comprises a time-series with 1000 observations, not 100. Consequently, the level of certainty one has about the final model results—for REC models, or indeed any time-series analysis of radiocarbon-dated events—should take into account the sampling rate with respect to the number of dates and not the resolution/length of the time-series those dates have been converted into. To some extent, sampling the date densities accounts for this problem in REC models because, as explained earlier, even with two nearly overlapping date densities co-occurrence between sampled dates is very unlikely (Fig. 3C). Thus, chronological uncertainty is likely to erase signals given low sample sizes. Nevertheless, sample size needs careful consideration and large samples (with high sample intensity per unit space/time) are probably required.

It should be stressed here, too, that other well-known sampling issues affect REC models. As with any quantitative analysis of radiocarbon dates, samples are likely to be spatio-temporally non-independent, leading potentially to autocorrelation in terms of both sample date and collection intensity—archaeologists, for example, may take multiple dates from individual depositional contexts and large archaeological sites may attract more intense sampling. Analogous problems may arise in palaeoenvironmental research. Another potential problem concerns sample heterogeneity. Scholars will need to consider whether all the samples in a given database are measuring the same thing; for example, do a human bone and a seed from a hearth both contribute equally to an estimate of past population size? The REC model does not directly account for these sampling issues and researchers need to carefully consider sampling strategies, the inferential logic of date-as-data with respect to a given target process, and ‘chronometric hygiene’ (see Williams, 2012, for further discussion).

Future development of REC models could proceed in multiple directions. One involves further statistical power analyses. This research would probably involve a comprehensive simulation study aimed at determining the true/false positive rate for REC models given a set of realistic parameters. Limits and ranges for those parameters (numbers of events, degree of chronological uncertainty, strength of the true effect, etc.) could be drawn from a review of published studies based on radiocarbon-date proxies. Importantly, this sort of research should include determining minimum sample density requirements given a variety of models and different research objectives. It should also involve specific attention to the impact of calibration curve uncertainty and, more specifically, how that may vary for different sections of a given calibration curve—see the Supporting Information for a brief exploration of the IntCal13 curve (Reimer *et al.*, 2013).

A second direction for future research should involve a deeper investigation into the magnitude of the bias caused by temporal spread and potential ways of reducing that bias. In the experiments described here, the bias seemed to consistently reduce the magnitude of the true parameter value by close to half. While it seems unlikely that the estimated magnitude would usually be half that of the true value, it may be that the magnitude of the bias is a function of some analytical or experimental condition, such as the average degree of chronological uncertainty among individual events, or the average slope of the calibration curve over the span of the true event-count sequence. For example, the extended analysis described in the Supporting Information involved a much earlier portion of the calibration curve (22 000 BP). That analysis produced a parameter estimate that biased further towards zero than the estimates described above, which suggests that there may be a relationship between curve location and the degree of bias. If that pattern holds and the bias was predictable given some other parameter or characteristic of the relevant data, a correction might

be applied. Such a correction, if it exists, would obviously be very useful. Other practical corrections could be explored as well, however, such as beginning with a set of dates covering a long interval and then carefully selecting a smaller window of time for analysis—although this could lead to a ‘garden of forking paths’ (Gelman and Loken, 2014) and trimming the data in this way should be done with caution.

Another direction for future research could involve improving the specification of model parameters. In particular, the p parameter in the negative binomial model used for the present study was left to be estimated from the data. This may ultimately be the best approach, but information from a given dataset could be used to constrain that parameter. The parameter could represent the probability that any events occurred at a given time, which suggests that an SPDF proxy or perhaps the KDE model could potentially be used to inform the parameter estimates.

A fourth direction for future work could involve comparing REC models to the other published approaches. The cursory exploration described above involving the SPDF-GLM and the REC model is only scratching the surface and more systematic comparisons should be conducted to fully understand the observable differences. A detailed comparison that includes mathematical details might also lead to a deeper understanding of the nature of chronological uncertainty and its impact on patterns observed with different regression models.

A final direction for future research should involve developing a way to incorporate information from Bayesian chronological models to constrain uncertainty. In the present study, the events were each dated with a single simulated radiocarbon date density. This was analogous to a radiocarbon-date database consisting of stratigraphically unrelated events. Fully specified chronological models could be used, however, if the events were stratigraphically related. In such cases, simply sampling the marginal posterior calibrated radiocarbon-date densities—as was done here—would not be taking advantage of the available information. Instead, a chronological model should be used to constrain the draws from the set of probable event-count sequences so that they respect prior information about the temporal relationships among events. Recent developments involving the application of Bayesian hierarchical modelling to chronological analyses of archaeologically defined cultural phases could potentially be useful in this regard (Banks *et al.*, 2019). These Bayesian hierarchical phase models might be used to constrain the sampling of a given RECE in order to incorporate both stratigraphic and archaeological prior knowledge into REC models.

Conclusions

REC models appear to be a useful way to extract information from large databases of radiocarbon-dated events. Unlike analyses involving established summary proxies based on radiocarbon dates—e.g., the well-known SPDF—REC models do not conflate process variation with chronological uncertainty. Instead, REC models account for the impact of that uncertainty on parameter estimates by sampling the probable arrangements of event-count sequences given the chronological uncertainty in radiocarbon dates. REC model parameter estimates have signs that reflect the true process–covariate relationship and are biased towards zero. While the bias needs to be considered when conducting research in practice, it appears to be consistent (always towards zero), which means that the models will be useful in an applied context. REC models are, therefore, a promising alternative to established approaches.

Supporting information

Additional supporting information may be found in the online version of this article at the publisher's web-site.

Acknowledgements. While no special funding was used to perform this research, employment with the independent Extreme Events Research Group at the Max Planck Society made it possible. The group is affiliated with and financially supported by the Max Planck Institutes for Chemical Ecology, The Science of Human History, and Biogeochemistry, all of which are located in Jena, Germany. I would also like to thank the Editor-in-Chief of the *Journal of Quaternary Science*, Prof. C. Neil Roberts, and two reviewers for providing helpful feedback that improved the paper. Open access funding enabled and organized by Projekt DEAL.

Data availability statement

Data used in this study were primary simulated and can be reproduced by following the instructions in the 'Replication Document' included as a supplementary file with this paper, or available at www.github.com/wcarleton/recm. The palaeoclimate time series used in some simulations was downloaded from the NOAA data repository, which can be found at <https://www.ncdc.noaa.gov/paleo-search/study/13519> or in the same GitHub repository as the replication document and R code used in this study.

Abbreviations. CKDE, composite kernel density estimation; GLM, generalized linear model; KDE, kernel density estimation; MCMC, Markov chain Monte Carlo; REC, radiocarbon-dated event count; RECE, radiocarbon-dated event-count ensemble; SPDF, summed probability density function.

References

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

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

Banks WE, Bertran P, Ducasse S *et al.* 2019. An application of hierarchical Bayesian modeling to better constrain the chronologies of Upper Paleolithic archaeological cultures in France between ca. 32,000–21,000 calibrated years before present. *Quaternary Science Reviews* **220**: 188–214.

Blaauw M. 2020. clam: Classical Age-Depth Modelling of Cores from Deposits. R package version 2.3.4. <https://cran.r-project.org/web/packages/clam/index.html>

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

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

Bronk Ramsey CB. 2017. Methods for summarizing radiocarbon datasets. *Radiocarbon* **59**: 1809–1833.

Broughton JM, Weitzel EM. 2018. Population reconstructions for humans and megafauna suggest mixed causes for North American Pleistocene extinctions. *Nature Communications* **9**: 5441.

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.

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

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

Crema ER, Bevan A, 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 ^{14}C dates suggests regional divergences in the population dynamics of the Jomon period in Eastern Japan. *PLoS ONE* **11**: e0154809.

Crema ER, Kobayashi K. 2020. A multi-proxy inference of Jōmon population dynamics using Bayesian phase models, residential data, and summed probability distribution of ^{14}C dates. *Journal of Archaeological Science*, **117**: 105136. <http://dx.doi.org/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 of the United States of America* **113**: 14483–14491.

Dye T, 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.

Edinburgh 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**: 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**: 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**: 193–223.

Gelman A, Carlin JB, Stern HS *et al.* 2013. *Bayesian Data Analysis*. CRC Press Press: Boca Raton.

Gelman A, Hill J. 2007. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press: Cambridge.

Gelman A, Loken E. 2014. The statistical crisis in science. *American Scientist* **102**: 460.

Geweke J. 1992. Evaluating the accuracy of sampling-based Approaches to the calculation of posterior moments. In *Bayesian Statistics*, Bernardo J, Berger J, Dawid A, Smith A (eds). Clarendon Press: Oxford; 169–193.

Geyh MA. 1969. Versuch einer chronologischen Gliederung Des marinen Holozäns an der Nordseeküste mit Hilfe der statistischen Auswertung von ^{14}C -Daten. *Zeitschrift der Deutschen Geologischen Gesellschaft* **118**: 351–360.

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

Hilbe JM. 2011. *Negative Binomial Regression*. Cambridge University Press: Cambridge.

Hilbe JM. 2014. *Modeling Count Data*. Cambridge University Press: Cambridge.

Hoffmann T, Lang A, Dikau R. 2008. Holocene river activity: analysing ^{14}C -dated fluvial and colluvial sediments from Germany. *Quaternary Science Reviews* **27**: 2031–2040.

Kassambara A. 2019. *ggpubr: 'ggplot2' Based Publication Ready Plots*. R package version 0.2.2. <https://cran.r-project.org/web/packages/ggpubr/index.html>

Kennett DJ, Breitenbach SFM, Aquino VV *et al.* 2012. Development and disintegration of Maya political systems in response to climate change. *Science* **338**: 788–791.

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

Leipe C, Long T, Sergusheva EA *et al.* 2019. Discontinuous spread of millet agriculture in eastern Asia and prehistoric population dynamics. *Science Advances* **5**: 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**: 267–293.

- Manning K, Timpson A. 2014. The demographic response to Holocene climate change in the Sahara. *Quaternary Science Reviews* **101**: 28–35.
- McLaughlin R, Hannah E, 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* **9**: 1–24.
- McLaughlin TR. 2019. On applications of space–time modelling with open-source ¹⁴C Age calibration. *Journal of Archaeological Method and Theory* **26**: 479–501.
- Mooney SD, Harrison SP, Bartlein PJ *et al.* 2011. Late Quaternary fire regimes of Australasia. *Quaternary Science Reviews* **30**: 28–46.
- NIMBLE Development Team. 2018. Nimble user manual. *R package manual version 0*: 6–12. <https://r-nimble.org/>
- Pierce JL, Meyer GA, Jull AJT. 2004. Fire-induced erosion and millennial-scale climate change in northern ponderosa pine forests. *Nature* **432**: 87–90.
- Prentiss AM, Cail HS, Smith LM. 2014. At the Malthusian ceiling: subsistence and inequality at Bridge River, British Columbia. *Journal of Anthropological Archaeology* **33**: 34–48.
- R Core Team. 2019. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing: Vienna. <https://www.r-project.org/>
- Ramsey JB. 1969. Tests for specification errors in classical linear least-squares regression analysis. *Journal of the Royal Statistical Society: Series B (Methodological)* **31**: 350–371.
- Reimer PJ, Bard E, Bayliss A *et al.* 2013. IntCal13 and Marine13 radiocarbon age calibration curves 0–50,000 years cal BP. *Radiocarbon* **55**: 1869–1887.
- Rick JW. 1987. Dates as data: an examination of the Peruvian Preceramic radiocarbon record. *American Antiquity* **52**: 55–73.
- Riede F. 2009. Climate and demography in early prehistory: using calibrated ¹⁴C dates as population proxies. *Human Biology* **81**: 309–337.
- Schulting R. 2010. Holocene environmental change and the Mesolithic-Neolithic transition in north-west Europe: revisiting twomodels. *Environmental Archaeology* **15**: 160–172.
- Shennan S. 2009. Evolutionary demography and the population history of the European early Neolithic. *Human Biology* **81**: 339–355.
- Shennan S. 2013. Demographic continuities and discontinuities in Neolithic Europe: evidence, methods and implications. *Journal of Archaeological Method and Theory* **20**: 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**: 2486.
- Steele J. 2010. Radiocarbon dates as data: quantitative strategies for estimating colonization front speeds and event densities. *Journal of Archaeological Science* **37**: 2017–2030.
- Taylor RE, Bar-Yosef O, Renfrew C. 2014. *Radiocarbon Dating. An Archaeological Perspective*. Albuquerque: Routledge.
- Thorndycraft VR, Benito G. 2006. The Holocene fluvial chronology of Spain: evidence from a newly compiled radiocarbon database. *Quaternary Science Reviews* **25**: 223–234.
- Turney CSM, Brown H. 2007. Catastrophic Early Holocene sea level rise, human migration and the Neolithic transition in Europe. *Quaternary Science Reviews* **26**: 2036–2041.
- Wickham H. 2016. *ggplot2: Elegant Graphics for Data Analysis*. Springer: New York. <https://cran.r-project.org/web/packages/ggpubr/index.html>
- Williams AN. 2012. The use of summed radiocarbon probability distributions in archaeology: a review of methods. *Journal of Archaeological Science* **39**: 578–589.
- Zhu F. 2011. A negative binomial integer-valued GARCH model. *Journal of Time Series Analysis* **32**: 54–67.