RADIOCARBON

CAMBRIDGE UNIVERSITY PRESS

# Inference from large sets of radiocarbon dates: software and methods

| | |
|---|---|
| Journal: | *Radiocarbon* |
| Manuscript ID | RDC-RES-2019-0037.R3 |
| Manuscript Type: | Research Article |
| Date Submitted by the Author: | 18-Jun-2020 |
| Complete List of Authors: | Crema, Enrico Ryunosuke; University of Cambridge, Archaeology<br>Bevan, Andrew; UCL, Institute of Archaeology |
| Keywords: | Radiocarbon dating, Summed Probability Distribution, Demography, Statistical analysis, Open Source Software |
| Abstract: | The last decade has seen the development of a range of new statistical and computational techniques for analysing large collections of radiocarbon dates, often but not exclusively to make inferences about human population change in the past. Here we introduce rcarbon, an open-source software package for the R statistical computing language which implements many of these techniques and looks to foster transparent future study of their strengths and weaknesses. In this paper, we review the key assumptions, limitations and potentials behind statistical analyses of summed probability distribution of radiocarbon dates, including Monte-Carlo simulation-based tests, permutation tests, and spatial analyses. Supplementary material provides a fully reproducible analysis with further details not covered in the main paper. |
| | |

# Inference from large sets of radiocarbon dates: software and methods

**Abstract**

The last decade has seen the development of a range of new statistical and computational techniques for analysing large collections of radiocarbon dates, often but not exclusively to make inferences about human population change in the past. Here we introduce *rcarbon*, an open-source software package for the R statistical computing language which implements many of these techniques and looks to foster transparent future study of their strengths and weaknesses. In this paper, we review the key assumptions, limitations and potentials behind statistical analyses of summed probability distribution of radiocarbon dates, , including Monte-Carlo simulation-based tests, permutation tests, and spatial analyses. Supplementary material provides a fully reproducible analysis with further details not covered in the main paper.

## 1. Introduction

The last few years has seen a dramatic increase in the number of research projects constructing proxy time series of demographic change out of large lists of archaeological radiocarbon dates. Put simply, this approach assumes that, given a large enough set of radiocarbon dates taken on anthropogenic samples, then the changing frequency of dates through time will preserve a signal of highs and lows in past human activity and, by extension, in human population. Rick's (1987) work was pioneering in this regard, being the first to propose the key assumption that more people in a given chronological period would typically lead to more anthropogenic products entering the archaeological record in that period, implying more potential samples to date and ultimately more published radiocarbon dates. He also already noted the presence of biases that were likely to distort such a signal (1987: fig.1). While early experiments with such methods sometimes considered a histogram of uncalibrated conventional radiocarbon ages, researchers have since turned to the summation of the posterior probability distributions of calibrated dates, and the result has become commonly known as a summed probability distribution (hereafter SPD, although there have also been alternative names and formulations).

The sharply increasing popularity of SPDs over the last decade or so has rightly also prompted criticism, not only with regard to the overall inferential assumptions behind the idea, but also with respect to the viability of particular SPD-based analytical methods. For example, several researchers have emphasised the fact that the sampling intensity of radiocarbon dates might not be constant over time. A good example is the difference between the popularity of radiocarbon sampling in early Mediterranean prehistory (e.g. Mesolithic-Neolithic) versus its almost complete avoidance for the Greek or Roman periods of the same region, even though the latter was manifestly a period of considerable population (Palmisano et al 2017). In addition to the impact of this differing prioritisation of absolute versus relative dating by archaeologists working on different time periods, researchers have further suggested that different kinds of societies (of otherwise roughly similar population size, for instance) might conceivably produce different radiocarbon footprints and/or that, even if a correlation between dates and population exists, that these might not scale in a linear fashion (Freeman et al. 2017). Others have noted that there might be a taphonomic bias towards the preservation of more anthropogenic material from sites of later periods (Surovell and Brantingham 2007; Surovell et al. 2009), again implying that over extended periods of thousands of years, we should probably assume a non-linear scaling to human activity. Such critiques are often valid to some degree and focus on how we should interpret summed probability distributions of radiocarbon dates in the first place (see discussions in Contreras and

Meadows 2014; Mokkonen 2014; Tallavara et al 2014; Attenbrow and Hiscock 2015; Hiscock and Attenbrow 2016; Smith 2016; Williams and Ulm 2016) Indeed, some of these very same issues also apply to other attempts to reconstruct past population (e.g. settlement counts where again it is sometimes difficult to compare evenly across periods and regions).

SPDs however also face a further challenge at a more fundamental level with regard to how best we might measure the changing frequencies of radiocarbon dates through time. Because calibrated radiocarbon dates comprise probability distributions spread across multiple calendar years and not discrete single estimates, the visual interpretation of aggregated SPDs becomes challenging and very often misleading at multiple scales. Peaks and troughs in SPDs might reflect changes in date intensity through time (and hence interpreted as population 'booms' or 'busts'), but they might also be a consequence of the changing steepness of the calibration curve, the size of the dates' associated measurement errors and/or just a statistical fluke from small sample sizes. In response to these challenges, a number of studies (Shennan and Edinborough 2007; Shennan et al. 2013; Timpson et al. 2014; Crema et al. 2016; Bevan et al 2017; Bronk Ramsey 2017; Brown 2017; Crema et al. 2017; Edinborough et al. 2017; Freeman et al. 2018; McLaughlin 2018; Roberts et al. 2018) have developed new techniques to address some of these issues. Most notably, they have offered new approaches to the problem of discerning genuine fluctuations in the density of radiocarbon dates as opposed to statistical artefacts arising from sampling error, the calibration process or taphonomic histories. Even so, replication and reuse of such methods remains limited, due both to an understandable experimentation across multiple software packages for calibration and statistical analysis (e.g. OxCal, CalPal, and in various forms via the R statistical environment, see **Supplementary Figure 1**) and to only patchy provision, so far, of transparent and reproducible workflows.

With a view to exploring and alleviating some of these issues, as well as with an eye to an increasing emphasis across archaeology and many other subjects on reproducible research (see Marwick 2016; Marwick et al 2017), we have recently developed *rcarbon* as an extension package for *R* (R Core Team 2018), one of the most popular software environments for statistical computing. The *rcarbon* package provides basic calibration, aggregation, and visualisation functions comparable to those that exist in other software packages, but also offers a suite of further functions for simulation-based statistical analysis of SPDs. This paper will discuss the main features of *rcarbon*, will highlight technical details and their implications in the creation and analyses of SPDs, and will offer some additional thoughts on the strengths and weakness of SPD-based methods overall.[1]

## 2. Calibration and Aggregation

### *2.1 Basic Treatment: Calibration and Summation*

In its most basic form an SPD extends the idea of a plotting a simple histogram of either uncalibrated $^{14}C$ ages or median calibrated dates to represent changing density of radiocarbon samples over time. Hence, the construction of an SPD involves two steps: (1) radiocarbon dates are calibrated so that for each sample we obtain a distribution of probabilities that the sample in question belongs to a particular

[1] Readers interested in applying these techniques on their own data are encouraged to read the R vignette associated with the package (https://cran.r-project.org/web/packages/rcarbon/vignettes/rcarbon.html). The supplementary material contains additional commentary and scripts for reproducing the analysis in the main paper. A copy of the supplementary material can also be accessed from the following repository: https://github.com/ercrema/rcarbon_paper_esm.

calendar year; and (2) all of these per-year probabilities are summed.[2] The resulting curve thus no longer represents probabilities, but instead is taken as a measure of date intensity. The rationale is thus not dissimilar to intensity-based techniques such as a univariate kernel density estimate (KDE), although with a crucial difference. In the case of KDE, individual kernels associated to each sample have all the same shape defined by the kernel bandwidth, itself mathematically estimated. In contrast, in the case of SPDs, the probability distributions associated with each radiocarbon date have different shapes depending on measurement error and the particularities of the relevant portion of the calibration curve. Consequently, SPDs are not explicitly and straightforwardly an estimate of the underlying distribution from which the observations are sampled from, and its absolute values cannot be directly compared across datasets. It follows that their visual interpretation within and across datasets is intrinsically biased.

Basic calibration in *rcarbon* is conducted with reference either to one of the established marine or terrestrial calibration curves or to a user-specific custom curve (in what follows, IntCal13 is used throughout: Reimer et al. 2013). The arithmetic method is for all intents and purposes identical to the the one adopted by OxCal (Bronk Ramsay 2008; leaving aside for a moment the more sophisticated Bayesian routines the latter package uses for more complex phase modelling), and very similar to that used by most other calibration software (Weninger et al 2015; Parnell 2018). Some of the terminology used by *rcarbon*'s standard routine has also been made consistent with *Bchron*, a well-known R package for handling radiocarbon dates and modelling pollen core chronologies and other age-depth relationships (Haslett and Parnell 2008; Parnell 2018; see also the *clam* package; Blaaw 2019). In *rcarbon*, the raw data stored for any given calibrated date consists of probability values per calibrated calendar year BP (but convertible to other calendars such as BC/AD), and it is these per-year probabilities that get summed to produce an SPD. For example, **Figure 1a** shows the result of adding up 130 dates from the Neolithic flint mines of Grimes Graves, Norfolk with three individual dates shown on top (for a full set and and more recent dates from the site, see Healy et al. 2014). A final point to note is that many studies apply a final 'smoothing function' to the SPD (e.g. Kelly et al 2013,Timpson et al 2014, Crema et al 2016, etc.), such as a running mean of between 50 and 200 years, to limit possible artefacts resulting from sampling error (but also from the effects of the calibration process) and discourage over-interpretation of the results (in **Figure 1a** an example with a 50-year running mean is shown). We return to the pros and cons of such smoothing in what follows.

*2.2 Phase or Site Over-Representation: Thinning and Binning*

In most instances, rather than the single site example provided above, an SPD is constructed across a wider region and using more than one site. As a result, there are further potential biases arising from the fact that not all sites (or indeed site phases) may have received equivalent levels of investment in radiocarbon dating. The Neolithic flint mining site of Grimes Graves in south-eastern England, for instance, has received an unusual level of investment in dating compared to other British prehistoric sites, but such differences do not accurately reflect a site's relative size or longevity of use. The cumulative effect of these differences in inter-site sampling intensity, and in particular the presence of abnormally high levels of sampling intensity of particular contexts, could thus generate artificial signals in the SPD. While the ideal approach to the problem is to select only samples referring to specific types of events (e.g. the construction of residential features) and control for sampling intensity via Bayesian inference (e.g. using OxCal's *R_Combine* function), the use of larger datasets with heterogeneous samples makes this solution unfeasible.

[2] In some software (e.g. CalPal), these two steps can be reversed (uncalibrated dates are summed and then the resulting aggregate is calibrated in one go), and we discuss the implications of this further below.

There are two alternative approaches to account for heterogeneity in sampling intensity. The first one involves manually going through a list of radiocarbon dates and choosing only a maximum number of better (e.g. short-lived, low-error) dates per phase or per site. In *rcarbon*, this thinning approach can also be achieved (in a less attentive but more automatic manner) using the *thinDates* function which either selects a maximum subset of dates at random or with a mixed approach that allows for some prioritisation of dates with lower errors (**Figure 1b**). This approach effectively replaces a set of radiocarbon dates referring to the same "event" with a smaller subset with user-defined size and inclusion criteria. As a consequence, the potentially biased contribution to the SPD of events associated with a larger number of radiocarbon dates can be reduced. A second solution to reduce the potential effect of such bias is to aggregate samples from the same site that are close in time, sum their probabilities, and divide the resulting SPD by the number of dates. Such site or phase-level 'binning' was introduced by Shennan et al. (2013) and discussed in detail by Timpson et al. (2014). The rationale is effectively to generate a local SPD referring to a particular occupation phase and to normalise this curve to unity to reduce the impact of heterogeneous sampling intensity. The *rcarbon* package provides a routine (*binPrep*), similar but not identical to the ones used in those two discussions, whereby dates from the same sites are grouped based on their (uncalibrated or median calibrated) inter-distances in time, defined by the parameter *h*, and then put into bins. Dates within the same bins are then aggregated to produce a local SPD that is normalised to sum to unity before being aggregated with other dates (and local SPDs) to produce the final curve. .

Different authors have already used different values for *h* (or comparable parameters) ranging anywhere from 50 to200 years (e.g. Shennan et al. 2013; Timpson et al. 2014; Crema et al. 2016; Bevan et al. 2017; Roberts et al. 2018). These choices can have a considerable effect on the resulting shape of the within site or within-phase local SPD, with higher values effectively leading to a more spread-out distribution of probabilities (**Figures 1c-e**) and we recommend exploring the implications of this empirically (e.g. via the *binSense* routine in *rcarbon* package (see also Riris 2018). It is also worth noting that there has been little or no discussion on what exactly constitutes a *bin* (or the "event" on which the thinning procedure is based), and how this might differ as a function of *h*, and ultimately affect the interpretation of SPDs. For example, *bins* generated from larger values of *h* effectively lead to an equal contribution of (potentially differently sized) sites to the SPD, effectively making this a proxy of site density rather than population size.

[ Figure 1 Here ]

*Figure 1. Summing, thinning and binning: (a) a summed probability distribution of dates from one site only (n=130 dates), with a slightly smoothed version also shown, as well as three example dates, followed by comparison of the smoothed raw density with (b) a randomly 'thinned' dataset of just 10 dates from the same site, (c-e) binned datasets at clustering cut-offs of h=50, 100 and 200 respectively.*

### *2.3 Normalised vs Unnormalised dates*

It is well-known that the shapes of individual calibrated probability distributions vary depending on the steepness or flatness of the calibration curve at that point in time. Less well-known is the fact that the area-under-the-curve of a date, calibrated in the usual arithmetic way, will not immediately sum to unity, but instead is typically normalised to ensure that it does (i.e. by dividing by the total sum under the curve for that date). **Figures 2a-b** provide two examples of dates at flat and steep portions of the calibration curve respectively which produce dramatically different areas-under-the-curve before normalisation. Weninger et al. (2015) first noted that the presence of this normalising correction explains the 'artificial spikes' noted by several different studies of SPDs, in which such spikes occurred in predictable ways at steep portions of the calibration curve (and which sometimes prompted attempts to smooth them away via fairly aggressive moving averages and/or various forms of kernel density estimate (see Williams 2012; Shennan et al. 2013; Timpson et al. 2014; Brown 2015, 2017; Ramsey 2017; McLaughlin 2018). **Figures 2c-e** provide three globally wide-ranging examples from the literature of datasets where spikes have been observed, with those spikes being particularly pronounced in early Holocene time series. In contrast, when unnormalised dates are summed, such spikes are not present. On first consideration, it is tempting to deem the normalised dates more theoretically justifiable, regardless of the spikes, because each date is seemingly 'treated equally' (i.e. each has a weight of 1 in the summation). However, because the summing a set of unnormalised calibrated dates (with varying post calibration areas under the curve) produces exactly the same result as first summing a set of uncalibrated Gaussians conventional radiocarbon age distributions (each of unity weight) and then calibrating them in one go (the process in *CalPal*, and also achievable in *rcarbon*, although not the default: see **Supplementary Figure 2**), this theoretical premise of the 'equal treatment' of each sample (i.e. the issue of unnormalised dates yielding an area under the curve equal to unity) can in fact be argued both ways (see Weninger et al. 2015 for extensive discussion). Regardless, these issues urge a basic caution not to over-interpret SPD results without considerable attention to how individual highs and lows in the data may have arisen.

[Figure 2 Here]

*Figure 2. Comparisons of unnormalised and normalised dates and their consequences: (a) a single date at a flat portion of the calibration curve (area under the probability histogram: 1.337), (b) a single date at a steep portion of the calibration curve (area under the probability histogram: 0.452), (c) Southern Levantine SPD ($n_{dates}$= 657, $n_{sites}$= 119, $n_{bins}$= 413 ; data from Roberts et al 2018), (d) Sahara SPD ($n_{dates}$= 643, $n_{sites}$= 233, $n_{bins}$= 551 ; data from Manning and Timpson 2014), and (e) Brazil SPD ($n_{dates}$= 173, $n_{sites}$= 97, $n_{bins}$= 171 ; data from Bueno et al 2013).The orange bar highlights time-intervals associated with steeper portions of the calibration curve.*

## 3. Statistical Testing

While it is tempting to treat the SPD itself as an unproblematic end goal with which to make interpretations about past population dynamics, this is rarely true, and it is almost always important to pay additional analytical attention to a host of uncertainties that come with it. For example, aside from the concerns often voiced about whether the density of radiocarbon dates can be regarded as a reliable proxy (see above), it is also worth noting at least two more issues. First, an ordinary SPD does not depict the uncertainty associated with the fact that certain calendar years are more likely to accrue a more narrowly defined dated sample than others (see **Supplementary Figure 3** for a worked through example). Nor does it depict the further uncertainty associated with larger or smaller sample sizes of dates or their measurement errors. A large number of radiocarbon dates for a given study may well improve the chance of a good signal, but there is no magic threshold, as this depends very much on the scope and goals of the analysis (e.g. inferences about multi-millennial trends versus those about sub-millennial trends, inferences about perceived growth rates through time or instead about regional differences across geographic space).

### *3.1 Model Fitting and Hypothesis Testing*

There have been various attempts so far to address these uncertainties, most of them leveraging the flexibility of Monte Carlo-type conditional simulation in some fashion, although more formally Bayesian models have also been proposed (see final section). Perhaps the most well-known approach was introduced by Shennan et al (2013) and compares an observed SPD with a theoretical null hypothesis of population change, where the latter might for instance imply stability (e.g. a flat, uniform theoretical SPD), growth (e.g. an exponential theoretical model) or initial growth-and-plateau (e.g. a logistic model) to name just a few of the most common (e.g. Shennan et al 2013; Crema et al. 2016; Bevan et al. 2017, Fernández-López de Pablo et al 2019). The usual workflow involves (1) fitting such a theoretical model to the observed SPD, (2) drawing *s* dates proportional to the shape of this fitted model (where *s* matches the number of observed dates or the number of bins if the dates have been binned), (3) back-calibrating individual dates from calendar time to 14C age, and assigning an error to each by randomly sampling (with replacement) the observed 14C age errors in the input data, (4) generating a theoretical SPD from the simulated data obtained in steps 2 and 3(5) repeating steps 2-4 *n* times and generating a critical (e.g. 95%) envelope for the theoretical SPD given the sample size, and (6) computing the amount that the observed SPD falls outside the simulation envelope compared to the randomised runs to produce a global p-value (as extensively described by Timpson et al 2014). These general steps have separately implemented by several authors (Zahid et al. 2016; Crema et al. 2016; Porčić and Nikolić 2016; Silva and Vander Linden 2017) with some minor differences (e.g. the formula for calculating the p-value, screening for false positives, etc.), and effectively treats the observed SPD as something comparable to a test statistic.

This approach has had the great virtue of grappling with the uncertainties associated with SPDs directly, but it is worth noting nevertheless that the choice, fitting and simulation of a null model of this kind is not straightforward. First, there are non-trivial technical niceties to do with how such a model is fitted in terms of the error model (e.g. log-linear or non-linear), or the time interval over which the model is fitted versus the interval over which it is simulated (given that all SPDs suffer from edge effects at their start and end dates). Second and more importantly, a particular model of theoretical population change or stability has to be selected and justified on contextual grounds, with perhaps the idea of exponential growth carrying the most straightforward demographic assumptions (all other things being equal and in light of the very long-term trend towards higher global population densities that seems to support this), but with other models often providing better fit to data or allowing certain kinds of extrapolation (e.g. Silva and Vander Linden 2017). A final point to stress regards the general limitations associated with the whole null hypothesis-testing approach: with a large enough sample, it will always be possible to produce a 'significant' result, but this may not warrant the kind of interpretation archaeologists and others are often looking for (e.g. about population "booms" and "busts"). It is also worth noting that intervals identified as positive or negative deviations from the null model are based on the density of dates and not on the trajectory of growth or decline even though the latter may be more interpretatively relevant in many situations. This means that, for example, intervals with positive deviations might well include instances of a decline in the density of radiocarbon dates. The Monte-Carlo simulation framework can be easily adapted to take this into account, allowing for testing against growth rates (see **supplementary figure 4**). Finally, the 95% critical envelopes produced for assessments of localised departure of the observed SPD from a theoretical pattern or a second SPD (see below, figures 3-4 for examples) are indicators only and should not be read as a set of formal significance tests for all years as this runs the well-known risk of multiple testing (see Loosmore and Ford 2006: 1926, for similar issues associated with the Monte Carlo envelopes produced for spatial point pattern analysis).

Many existing implementations of this technique both fit and sample from their theoretical models in calendar time. A set of individual calendar years are first drawn proportional to the fitted model, then these are back-calibrated individually to become a set of conventional (uncalibrated) $^{14}C$ ages with small errors deriving from those associated with the calibration curve itself. Then, larger plausible error terms are added to mimic the instrumental measurement errors of the observed dates and each age (typically now a Gaussian probability distribution) is then calibrated back into calendar time before all of the simulated dates are then finally aggregated into an SPD. This procedure can be formally described by a marginal probability with the assumption of a discretized calendar timeline:

$$p(r) = \sum_t^T \Pr(t) \times p(r|\mu_t, \sigma_t^2) \qquad [1]$$

where *p(r)* is the probability of selecting a random sample with a $^{14}$C age *r*, Pr(*t*) is the probability obtained from the fitted theoretical model at the calendar year *t* within *T* points in time across the temporal window of analysis, $\mu_t$ and $\sigma_t$ are their corresponding date in $^{14}$C age and the associated error on the calibration curve, and $p(r|\mu_t, \sigma_t)$ refers to the Gaussian probability density function. Thus ,if we ignore binning, given an observed dataset with k radiocarbon dates and a theoretical model Pr(t), one could apply equation 1 to obtain k 14C ages, to which we can assign random instrumental measurement errors by resampling from the observed data.

The term Pr(*t*) is generally obtained by: 1) fitting a curve (via regression) to an observed SPD over a defined temporal window; and 2) transforming the fitted values (e.g. for each discrete calendar year) so they sum to unity. Shennan et al (2003) initially fitted an exponential curve (as a null expectation for population with a constant growth rate), but other models have also been applied subsequently (cf. Crema et al 2016, Bevan et al 2017). It is also worth noting that Pr(*t*) does not have to be based on observed SPDs, and could potentially be derived from theoretical expectations or other demographic proxies (see Crema and Kobayashi 2020 for an example).

The assumption behind this sampling and back-calibration procedure (referred to in *rcarbon* as the *calsample* method, due to its sampling in calendar time) is that it will directly emulate both the kinds of uncertainty associated with a given observed sample size, and the impact on an SPD of the non-linearities in the calibration curve itself. However, the relationship between calendar years and radiocarbon ages is not commutative in the way such an approach implies (in agreement with Weninger et al 2015), and major problems are encountered in certain narrow parts of the calendar timescale, coincident with the same zones of artificial spiking first described above. **Figures 3a-b** depict the problem for the later Pleistocene and earlier Holocene time-frame using the same dated as in **figure 2c**. As before, we can note the difference in terms of spiking observed at predictable portions of the calibration curve where such spikes are present if we normalise individual dates but absent if we do not. However, the simulated envelopes created by the *calsample* approach exhibit quite different statistical artefacts at these locations (slight, offset dips if dates are normalised and dramatic dips if they are not). In neither case, do they seem to emulate the observed patterns.



In contrast, one alternative for generating theoretical SPDs is to back-calibrate the entire fitted model in one go and then to weight the result *p(r)* by the expected probability of sampling *r* under a uniform model:

$$v(r) = \frac{\sum_t^T Pr(t|null) \times p(r|\mu_t, \sigma_t^2)}{\sum_t^T Pr(t|uniform) \times p(r|\mu_t, \sigma_t^2)} \qquad [2]$$

Here Pr(*t*|*null*) is the fitted model under the null hypothesis, and Pr(*t*|*uniform*) is the probabilities associated with a uniform distribution covering for the same temporal range *T*. *v(r)* is then normalised to unity:

$$w(r) = \frac{v(r)}{\sum_{r}^{R} v(r)} \qquad [3]$$

with $R$ being all the $^{14}C$ ages examined, most typically the range covered by the calibration curve.

Simulations following this approach then draw samples of uncalibrated ages from the back-calibrated model and calibrate these, before summing (this is therefore referred to in *rcarbon* as the *uncalsample* method, see also Roberts et al 2018; Bevan et al 2017 for applications). The adjustment of the probability of sampling specific $^{14}C$ ages according to a baseline uniform model allows for much better simulation of the presence and amplitude of artificial peaks in the SPD at steeper portions of the calibration curve when dates are normalised, and their absence when dates are left unnormalised (**Figures 3c-d**). However, we note that neither approach is likely to be ideal, and discuss some promising alternatives in the sections below.

[Figure 3 Here]

*Figure 3: The relationship between observed data and simulations envelopes for four different methods (using the same data as in figure 2c): calsample realisations of (a) normalised and (b) unnormalised dates, and uncalsample realisations of (c) normalised and (d) unnormalised dates. Temporal ranges highlighted in red and blue represent intervals where the observed SPD show a significant positive or negative deviation from the simulated envelope (they do not necessarily imply the onset point of significant growth or decline).*

### *3.2 Comparison and Testing of Multiple SPDs*

A key advantage of SPDs over more traditional proxies of prehistoric population change, such as settlement counts, is the greater ease with which trajectories across different geographical regions can be compared, without the analytically-awkward frameworks imposed by different relative artefact-based chronologies. With this in mind, Crema et al. (2016) developed a permutation-based test to statistically compare two or more SPDs. While the null hypothesis for the one-sample models discussed above is a user-supplied theoretical growth model (e.g. we should expect exponential population growth all other things being equal), the null hypothesis of the multi-sample approach is that the SPDs are samples derived from the same statistical population (e.g. there is no meaningful difference between the shape of the SPD for region A and the one for region B). As for the one-sample approach p-values are obtained via simulation, but in this case rather than generating samples from a theoretical fitted model, the label defining the membership of each date (or bin if binning is being used) is permuted (e.g. we shuffle which dates belong to group A and which ones belong to group B, then produce a new SPD for each group, and repeat many times). This approach can be used to compare SPDs from different regions (as in Crema et al. 2016; Bevan et al. 2017; Riris 2018; Roberts et al. 2018) in order to infer where local population dynamics differ significantly through time, but it can also be used to consider other groupings of dates, such as those taken on different kinds of physical radiocarbon sample (Bevan et al. 2017). Such a *mark permutation test* will generate simulation envelopes for each SPD whose width proportional to the sample size (i.e. the overall number of dates per region, or the overall number of bins if binning has been applied; **figure 4**). Similar to the case of the one-sample approach, both one global and a set of local p-values can be obtained, the former assessing whether there are significant overall differences between sets and the latter identifying particular portions of the SPD with important differences in the summed probabilities.

While there are certainly still ways to mis- or over-interpret the results of this kind of mark permutation test, one major strength is that they do not face quite the same problems associated with model selection, fitting and simulation that the one sample approach does.

[Figure 4 Here]

*Figure 4: Example of mark permutation test (Crema et al 2016), comparing the SPDs from Southern ($n_{dates}$= 657, $n_{sites}$= 119, $n_{bins}$= 413) and Northern Levant ( $n_{dates}$= 589, $n_{sites}$= 41, $n_{bins}$= 296). Temporal ranges highlighted in red and blue represents intervals where the observed SPD show a significant positive or negative deviation from the pan-regional null model. Data from Roberts et al 2018.*

## 4. Spatial Analysis

A regional mark permutation test such as described above already offers one way to compare different geographic regions, but its application requires a crisp definition of these regions from the outset and it is thus not a particularly flexible way to explore variation across continuously varying geographic spaces. Early extensions of the SPD approach already had further spatial inferences in mind when they made use of weighted kernel density estimates (KDE) to infer regions of high or low concentrations of dates across multiple temporal slices, occasionally using animations (e.g. Collard et al 2010; Manning and Timpson 2014). Such visual inspection can be the basis for developing specific hypotheses, but suffers from the same limitations as a non-spatial SPDs: it is hard to know what to interpret as interesting variation in date intensity, through time and space, versus variation introduced by the calibration process, by sampling error or by investigative bias. Recent spatio-temporal analyses of radiocarbon dates have tackled this issue in two distinct ways, and we consider each one in turn below.

### *4.2 Flexible Timeslice Mapping*

In *rcarbon*, for instance, it is possible to map the spatio-temporal intensity of observed radiocarbon dates as relevant for a particular `focal' year (using the *stkde* function). This is achieved by first computing weights associated with each sampling point $x$ given the `focal' year $f$ and temporal bandwidth $b$ using the following equation:

$$w(x,f,b) = \sum_i^T p_i(x) e^{\frac{-(i-f)^2}{2b^2}} \qquad [4]$$

where $p_i(x)$ is the probability mass associated with the year $i$ obtained from the calibration process. In other words, a temporal Gaussian kernel is placed around a chosen year and then the degree of overlap between this kernel and the probability distribution of each date is evaluated. Each georeferenced date also has a Gaussian distance-weighted influence on spatial intensity estimate at a given location on the map (with the help of the R package *spatstat*: Baddeley et al 2015): in other words, a spatio-temporal kernel is applied, with both the spatial and the temporal Gaussian bandwidths defined by the user. The choice of appropriate spatial and the temporal bandwidth can arise from data exploration which suggests combinations that are both empirically-useful (e.g. for the particular problem or question of interest) and practically-aware (e.g. of the positional and temporal uncertainties in the underlying data), or it can be made via one of several automatic bandwidth selectors (see Davies et al 2018 for a specific review tailored to spatio-temporal analysis). While the latter option has the advantage of avoiding somewhat arbitrary values for the kernel bandwidth, it is worth noting that the choice of different bandwidth selectors can lead to very different result, particularly in the context of spatio-temporal analysis where there is no single agreed algorithm[3]. **Figure 5a** shows an example of the resulting surface for the focal year 6000 calBP, while **figure 5b** shows an unchanging overall surface where all samples are treated equally regardless of their actual date (i.e. an ordinary kernel density map).

**Figure 5c** shows the result of dividing one by the other which offers an indication of the *proportion* of local dates belonging to the focal, target time period, thereby to some extent detrending for any recovery biases present in the overall sample. This is analogous and consistent with the idea of *relative risk* mapping (Kelsall and Diggle 1995; Bevan 2012) and such an approach has been used by Chaput et al (2015) and Bevan et al (2017) to investigate spatial variation in the radiocarbon density North America and in the British Isles respectively. **Figure 5d** shows a further and final useful measure is of 'change' between the focal year and some earlier reference or backsight year (e.g. 200 years before, with various options for how 'change' or growth/decline is expressed). Colour ramps can be standardised to allow comparison across time-slices and thus also animation through multiple timeslices.

[Figure 5 Here]

*Figure 5. Example output of one focal year of a kernel density map of English and Welsh dates from the Euroevol Neolithic dataset ($n_{dates}$= 2,327, $n_{sites}$= 653, $n_{bins}$= 1,461, data from Manning et al 2016): (a) the spatio-temporal intensity for the focal year 6000 calBP, (b) the overall spatial intensity for Neolithic dates (8000-4000 calBP), (c) the proportion of a) out of b), and (d) a measure of the spatial pattern of change, mostly growth, from 6200 calBP to 6000 calBP.*



### *4.2 Spatial Testing*

The above spatial mapping emphasises flexible visualisation, but a complementary second approach to spatial analysis or georeferenced radiocarbon lists instead prioritises the testing of any observed spatial trends, via an extension of the permutation method described above. It compares local SPDs (i.e. SPDs created at each observation point by weighting the radiocarbon contribution of neighbouring sites as a function of their distance to the focal point) to the expected local SPD under stationarity (i.e. all local SPD showing the same pattern), obtained via a random permutation of the spatial coordinates of each site. The result (**Figure 6**) provides a significance test for each site location, highlighting regions with higher or lower growth rates compared to the pan-regional trend (see also Crema et al 2017).

[Figure 6 Here]

*Figure 6. Spatial permutation test for the same data as figure 5 showing: (a) the local mean geometric growth rates mean geometric growth rate between 6300-6100 to 6100-5900 calBP; and (b) results of the spatial permutation test for the same interval showing local significant positive and negative significant departures from the null hypothesis.*

## 5. Conclusion

As the above should make clear, we continue to see great promise in the aggregate treatment of radiocarbon dates as proxies for activity intensity, and it is interesting to note that similar conclusions have been made in other fields that do not focus on human population, but instead use such lists to explore, amongst other things, alluvial accumulation, volcanic activity or peat deposition (Michczyńska and Pazdur 2004; Surovell et al. 2009; Macklin et al 2014). The basic notion behind an SPD remains relatively easy to understand and in part this is probably the reason for its widespread appeal, even if some of the ensuing testing methods become more complicated. The *rcarbon* package is an attempt to

[3] Users interested in applying these different bandwidth selectors are advised to consult the R packages *spatstat* (Baddeley et al 2015) and *sparr* (Davies et al 2018). For an archaeological review of univariate and bivariate bandwidth selectors see Baxter et al 1997. See also Bronk-Ramsay 2017 for an alternative approach to univariate KDE for radiocarbon dates.

provide a working environment within which to explore both the strengths and weaknesses of such an approach. There is also a useful transferability of SPD approaches to proxy time series constructed from other kinds of evidence, such as dendrochronological dates (Ljungqvist et al. 2018) or even traditionally-dated artefact datasets. Even so, there continues to be a real need to consider how alternatives, for example Gaussian mixtures (Parnell 2018), might offer superior and theoretically more coherent frameworks, and to grapple further with quantisation and calibration curve effects (Weninger and Clare 2018).

**Acknowledgments**

We would like to thank several colleagues with whom we have discussed many aspects of this paper, as well those who provided constructive feedback on the *rcarbon* package, in particular: Anna Bloxam, Kevan Edinborough, Alessio Palmisano, Peter Schauer, Stephen Shennan, Fabio Silva and Bernhard Weninger. We are also grateful to the three anonymous reviewers for their insightful comments on the manuscript. This project has received funding from the European Research Council (ERC) under the Horizon 2020 research and innovation programme (Grant Agreement No 801953). This material reflects only the authors' views and the Commission is not liable for any use that may be made of the information contained therein.



**References**

Attenbrow, V., Hiscock, P. 2015. Dates and demography: are radiometric dates a robust proxy for long-term prehistoric demographic change? *Archaeology in Oceania*, 50: 30–36.

Baddeley, A., Rubak, E., Turner, R. 2015. *Spatial Point Patterns: Methodology and Applications with R*. London: Chapman and Hall/CRC Press

Baxter,M.J.,Beardah, C.C., Wright, R.V.S. 1997. Some Archaeological Applications of Kernel Density Estimates. *Journal of Archaeological Science*, 24: 347-354.

Bevan, A. 2012. Spatial methods for analysing large-scale artefact inventories. *Antiquity*, 86(332): 492–506.

Bevan, A., Colledge, S., Fuller, D., Fyfe, R., Shennan, S., Stevens, C. 2017. Holocene fluctuations in human population demonstrate repeated links to food production and climate. *Proceedings of the National Academy of Sciences*, 114(49):E10524–E10531.

Blaauw M. 2019. clam: Classical Age-Depth Modelling of Cores from Deposits. R package version 2.3.2, URL: <https://CRAN.R-project.org/package=clam>

Bronk Ramsey, C. 2008. Radiocarbon dating: revolutions in understanding, *Archaeometry* 50(2): 249-75.

Bronk Ramsey, C. 2017. Methods for Summarizing Radiocarbon Datasets. *Radiocarbon*, 59(6): 1809–1833.

Brown, W. A. 2015. Through a filter, darkly: population size estimation, systematic error, and random error in radiocarbon-supported demographic temporal frequency analysis. *Journal of Archaeological Science*, 53: 133–147.

Brown, W. A. 2017. The past and future of growth rate estimation in demographic temporal frequency analysis: Biodemographic interpretability and the ascendance of dynamic growth models. *Journal of Archaeological Science*, 80: 96–108.

Chaput, M. A., Kriesche, B., Betts, M., Martindale, A., Kulik, R., Schmidt, V., Gajewski, K. 2015. Spatiotemporal distribution of Holocene populations in North America. *Proceedings of the National Academy of Sciences*, 112(39): 12127–12132.

Collard, M., Edinborough, K., Shennan, S.,Thomas, M. G. 2010. Radiocarbon evidence indicates that migrants introduced farming to Britain. *Journal of Archaeological Science,* 37(4): 866–870.

Contreras, D. A., Meadows, J. 2014. Summed radiocarbon calibrations as a population proxy: a critical evaluation using a realistic simulation approach. *Journal of Archaeological Science,* 52, 591–608.

Crema, E. R., Bevan, A., Shennan, S. 2017. Spatio-temporal approaches to archaeological radiocarbon dates. J*ournal of Archaeological Science:* 87, 1–9.

Crema, Enrico R., Habu, J., Kobayashi, K., Madella, M. 2016. Summed Probability Distribution of 14 C Dates Suggests Regional Divergences in the Population Dynamics of the Jomon Period in Eastern Japan. *PLOS ONE*, 11(4): e0154809. DOI:10.1371/journal.pone.0154809

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

Davies, Tilman M., Jonathan C. Marshall, and Martin L. Hazelton. (2018) Tutorial on Kernel Estimation of Continuous Spatial and Spatiotemporal Relative Risk. *Statistics in Medicine*, 37(7): 1191-1221

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

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

Freeman, J., Baggio, J. A., Robinson, E., Byers, D. A., Gayo, E., Finley, J. B., Meyer, J.A., Kelly, R.L., Anderies, J.M. 2018. Synchronization of energy consumption by human societies throughout the Holocene. *Proceedings of the National Academy of Sciences*, 115(40): 9962-9967.

Freeman, J., Byers, D. A., Robinson, E., Kelly, R. L. 2017. Culture Process and the Interpretation of Radiocarbon Data. *Radiocarbon*, 60(2): 453-467.

Haslett, J., Parnell, A.C., 2008. A simple monotone process with application to radiocarbon-dated depth chronologies, *Journal of the Royal Statistical Society: Series C Applied Statistics* 57.4: 399-418.

Hiscock, P., Attenbrow, V., 2016. Dates and demography? The need for caution in using radiometric dates as a robust proxy for prehistoric population change. *Archaeology in Oceania*, 51(3): 218–219.

Healy, F., Marshall, P., Bayliss, A., Cook, G., Bronk Ramsey, C., van der Plicht, J., Dunbar, E. 2014. *Grime's Graves, Weeting-with-Broomhill, Norfolk. Radiocarbon Dating and Chronological Modelling*, Portsmouth: Historic England Research Report 27/2014

Kelsall, J. E., Diggle, P. J. 1995. Non-parametric estimation of spatial variation in relative risk. *Statistics in Medicine,* 14(21–22): 2335–2342.

Ljungqvist, F.C., Tegel, W., Krusic, P.J., Seim, A., Gschwind, F.M., Haneca, K., Herzig, F., Heussner, K.-U., Hofmann, J., Houbrechts, D., Kontic, R., Kyncl, T., Leuschner, H.H., Nicolussi, K., Perrault, C., Pfeifer, K., Schmidhalter, M., Seifert, M., Walder, F., Westphal, T., Büntgen, U., 2018. Linking European building activity with plague history. *Journal of Archaeological Science* 98: 81–92.

Loosmore, N.B., Ford, E.D. 2006. Statistical inference using the G or K point pattern spatial statistics, *Ecology* 87: 1925-1931.

Macklin, M. G., Lewin, J., Jones, A.F. 2014. Anthropogenic alluvium: An evidence-based meta-analysis for the UK Holocene, *Anthropocene* 6: 26-38.

Marwick, B., 2017. Computational Reproducibility in Archaeological Research: Basic Principles and a Case Study of Their Implementation. *Journal of Archaeological Method and Theory,* 24: 424–450

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

Michczyńska, D., Pazdur, A. 2004. Shape Analysis of Cumulative Probability Density Function of Radiocarbon Dates Set in the Study of Climate Change in the Late Glacial and Holocene. *Radiocarbon* 46(2): 733-744.

McLaughlin, T. R. 2019. On Applications of Space–Time Modelling with Open-Source 14C Age Calibration. *Journal of Archaeological Method and Theory*, 26: 479-501.

Mökkönen, T. 2014. Archaeological radiocarbon dates as a population proxy: a skeptical view. *Fennosc. Archaeol.* 31: 125-134.

Parnell, A. 2018. Bchron: Radiocarbon Dating, Age-Depth Modelling, Relative Sea Level Rate Estimation, and Non-Parametric Phase Modelling, R package. URL: <https://CRAN.R-project.org/package=Bchron>

Porčić, M., Nikolić, M. 2016. The Approximate Bayesian Computation approach to reconstructing population dynamics and size from settlement data: demography of the Mesolithic-Neolithic transition at Lepenski Vir. *Archaeological and Anthropological Sciences* 8(1): 169–186.

R Core Team 2018. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: < https://www.R-project.org/.>

Reimer, P.J., Bard, E., Bayliss, A., Beck, J.W., Blackwell, P.G., Ramsey, C.B., Buck, C.E., Cheng, H., Edwards, R.L., Friedrich, M., Grootes, P.M., Guilderson, T.P., Haflidason, H., Hajdas, I., Hatté, C., Heaton, T.J., Hoffmann, D.L., Hogg, A.G., Hughen, K.A., Kaiser, K.F., Kromer, B., Manning, S.W., Niu, M., Reimer, R.W., Richards, D.A., Scott, E.M., Southon, J.R., Staff, R.A., Turney, C.S.M., Plicht, J. van der, 2013. IntCal13 and Marine13 Radiocarbon Age Calibration Curves 0–50,000 Years cal BP. *Radiocarbon* 55: 1869–1887.

Rick, J. W. 1987. Dates as Data: An Examination of the Peruvian Preceramic Radiocarbon Record. *American Antiquity*, 52: 55-73.

Riris, P. 2018. Dates as data revisited: A statistical examination of the Peruvian preceramic radiocarbon record. *Journal of Archaeological Science*, 97: 67–76.

Roberts, N., Woodbridge, J., Bevan, A., Palmisano, A., Shennan, S., Asouti, E. 2018. Human responses and non-responses to climatic variations during the last Glacial-Interglacial transition in the eastern Mediterranean. *Quaternary Science Reviews*, 184: 47–67.

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

Shennan, S., Edinborough, K. 2007. Prehistoric population history: from the Late Glacial to the Late Neolithic in Central and Northern Europe. Journal of Archaeological Science, 34, 1339–1345.
Silva, F., Vander Linden, M. 2017. Amplitude of travelling front as inferred from 14 C predicts levels of genetic admixture among European early farmers. *Scientific Reports*, 7(1): 11985. DOI:10.1038/s41598-017-12318-2

Smith, M. 2016. The use of summed-probability plots of radiocarbon data in archaeology. *Archaeology in Oceania,* 51(3): 214–215.

Surovell, T. A., Brantingham, P. J. 2007. A note on the use of temporal frequency distributions in studies of prehistoric demography. *Journal of Archaeological Science,* 34: 1868–1877.

Surovell, T.A., Byrd Finley, J., Smith, G. M., Brantingham, P.J., Kelly, R. 2009. Correcting temporal frequency distributions for taphonomic bias, *Journal of Archaeological Science,* 36: 1715–1724.

Tallavaara,M., Pesonen, P., Oinonen, M., Seppä, H. (2014). The mere possibility of biases does not invalidate archaeological population proxies–response to Teemu Mökkönen. *Fennosc. Archaeol*, 31: 135-140.

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

Williams, A. N., & Ulm, S. 2016. Radiometric dates are a robust proxy for long-term demographic change: A comment on Attenbrow and Hiscock (2015). *Archaeology in Oceania,* 51(3): 216–217.

Weninger, B., Clare, L., Jöris, O., Jung, R., Edinborough, K. 2015. Quantum theory of radiocarbon calibration. *World Archaeology*, 47(4): 543–566.

Weninger B. and Clare L. 2018. High-Resolution Chronology of Shir, South Area, In Bartl K (ed.). *The Late Neolithic Site of Shir/Syria. Volume I. The Excavations at the South Area 2006–2009. Damaszener Forschungen, Vol. 18. Archaölogische Forschungen in Syrie*n: 183-198. Darmstadt: Philipp von Zabern.

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

Zahid, H. J., Robinson, E., Kelly, R. L. 2016. Agriculture, population growth, and statistical analysis of the radiocarbon record. *Proceedings of the National Academy of Sciences*, 113(4): 931-935.

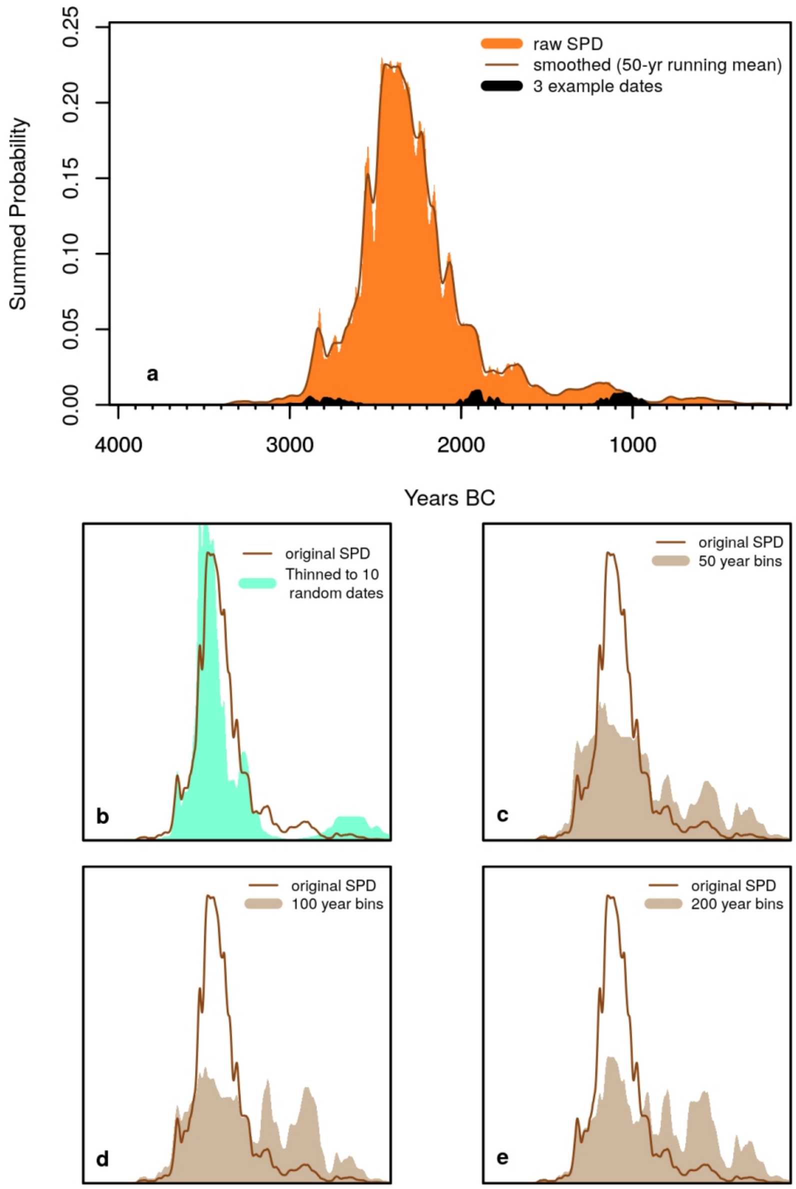


Figure 1. Summing, thinning and binning: (a) a summed probability distribution of dates from one site only (n=130 dates), with a slightly smoothed version also shown, as well as three example dates, followed by comparison of the smoothed raw density with (b) a randomly 'thinned' dataset of just 10 dates from the same site, (c-e) binned datasets at clustering cut-offs of h=50, 100 and 200 respectively.

101x152mm (300 x 300 DPI)

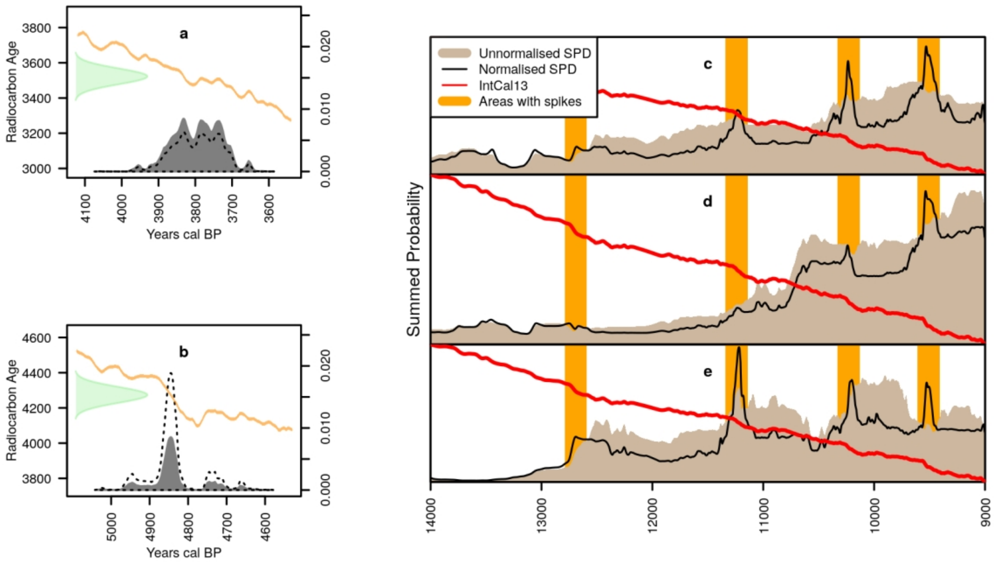


Figure 2. Comparisons of unnormalised and normalised dates and their consequences: (a) a single date at a flat portion of the calibration curve (area under the probability histogram: 1.337), (b) a single date at a steep portion of the calibration curve (area under the probability histogram: 0.452), (c) Southern Levantine SPD (ndates= 657, nsites= 119, nbins= 413 ; data from Roberts et al 2018), (d) Sahara SPD (ndates= 643, nsites= 233, nbins= 551 ; data from Manning and Timpson 2014), and (e) Brazil SPD (ndates= 173, nsites= 97, nbins= 171 ; data from Bueno et al 2013).The orange bar highlights time-intervals associated with steeper portions of the calibration curve.

139x88mm (300 x 300 DPI)

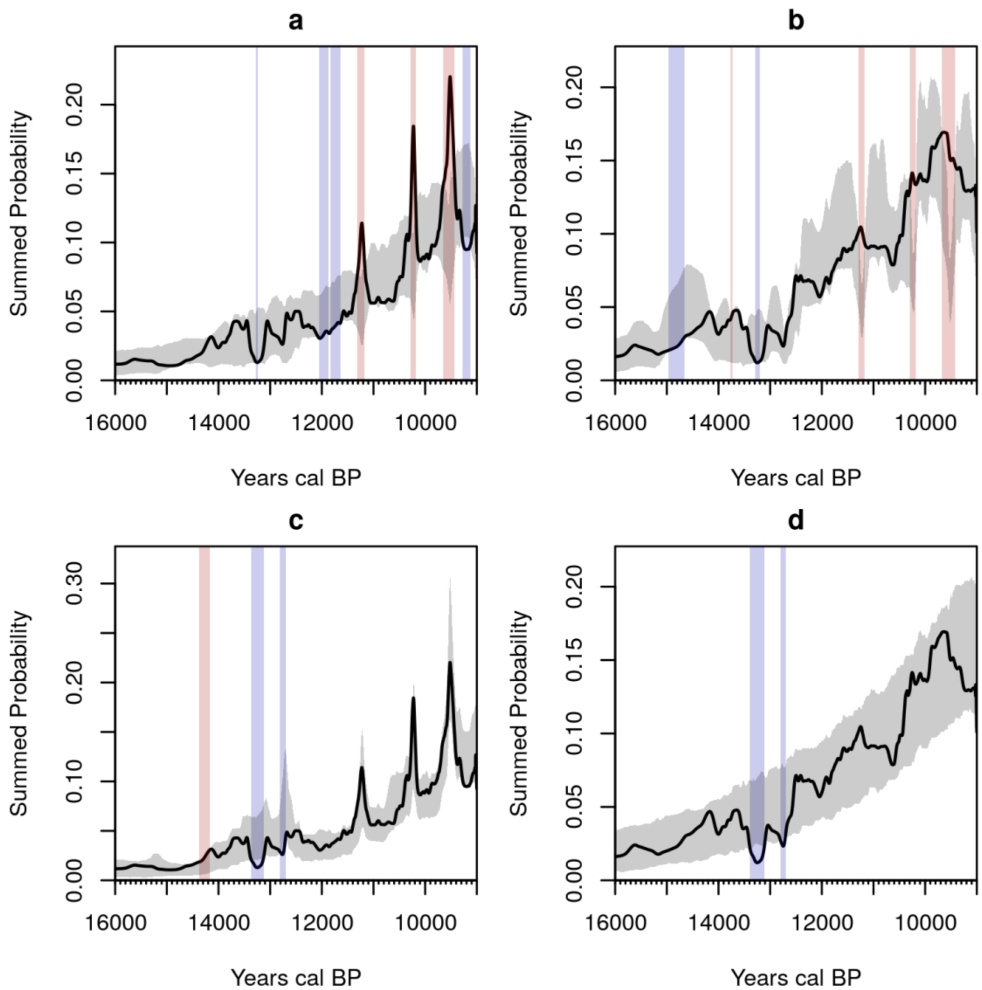


Figure 3: The relationship between observed data and simulations envelopes for four different methods (using the same data as in figure 2c): calsample realisations of (a) normalised and (b) unnormalised dates, and uncalsample realisations of (c) normalised and (d) unnormalised dates. Temporal ranges highlighted in red and blue represent intervals where the observed SPD show a significant positive or negative deviation from the simulated envelope (they do not necessarily imply the onset point of significant growth or decline).

127x127mm (300 x 300 DPI)

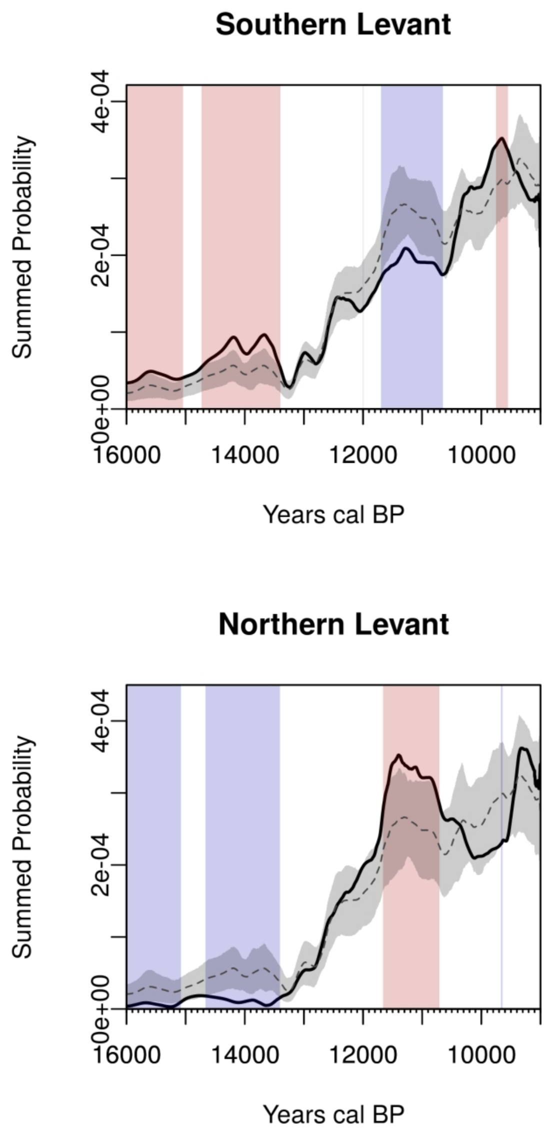


Figure 4: Example of mark permutation test (Crema et al 2016), comparing the SPDs from Southern (ndates= 657, nsites= 119, nbins= 413) and Northern Levant ( ndates= 589, nsites= 41, nbins= 296). Temporal ranges highlighted in red and blue represents intervals where the observed SPD show a significant positive or negative deviation from the pan-regional null model. Data from Roberts et al 2018.

101x203mm (300 x 300 DPI)

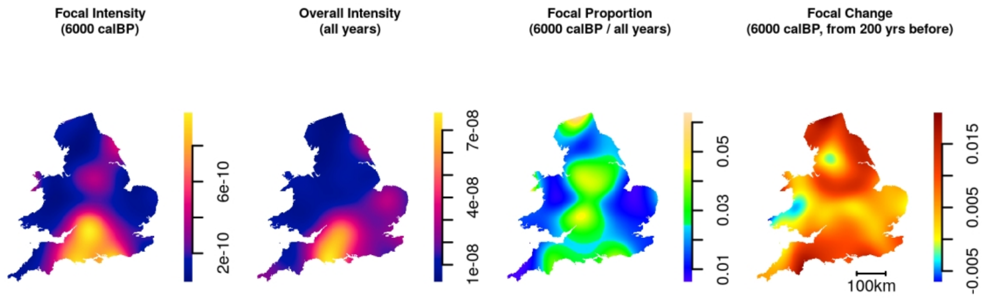


Figure 5. Example output of one focal year of a kernel density map of English and Welsh dates from the Euroevol Neolithic dataset (ndates= 2,327, nsites= 653, nbins= 1,461, data from Manning et al 2016): (a) the spatio-temporal intensity for the focal year 6000 calBP, (b) the overall spatial intensity for Neolithic dates (8000-4000 calBP), (c) the proportion of a) out of b), and (d) a measure of the spatial pattern of change, mostly growth, from 6200 calBP to 6000 calBP.

139x50mm (300 x 300 DPI)

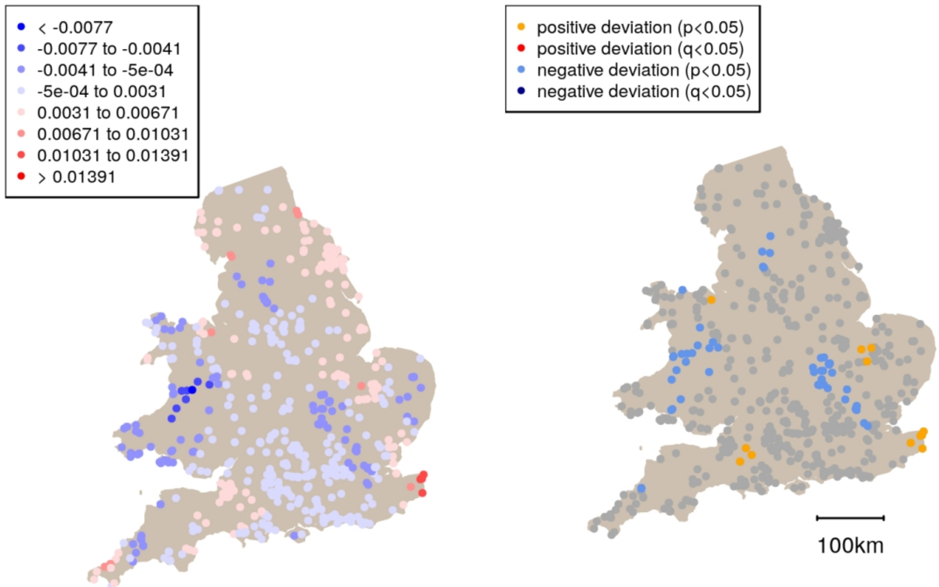


Figure 6. Spatial permutation test for the same data as figure 5 showing: (a) the local mean geometric growth rates mean geometric growth rate between 6300-6100 to 6100-5900 calBP; and (b) results of the spatial permutation test for the same interval showing local significant positive and negative significant departures from the null hypothesis.

127x88mm (300 x 300 DPI)