Journal of Archaeological Science 52 (2014) 591–608

Contents lists available at ScienceDirect

Journal of Archaeological Science

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

# Summed radiocarbon calibrations as a population proxy: a critical evaluation using a realistic simulation approach

Daniel A. Contreras [a, *], John Meadows [b, c]

[a] Institute for Ecosystem Research, Christian-Albrechts Universität zu Kiel, Olshausenstr. 75, D-24118 Kiel, Germany
[b] Zentrum für Baltische und Skandinavische Archäologie, Stiftung Schleswig-Holsteinische Landesmuseen, Schloss Gottorf, D-24837 Schleswig, Germany
[c] Christian-Albrechts-Universität zu Kiel, Leibniz-Labor für Altersbestimmung und Isotopenforschung, Max-Eyth-Str. 11-13, D-24118 Kiel, Germany

ARTICLE INFO

*Article history:*
Received 11 March 2014
Received in revised form
19 May 2014
Accepted 20 May 2014
Available online 9 June 2014

*Keywords:*
Prehistoric demography
Summed radiocarbon date distributions
Simulation
Calibration
Population proxies

ABSTRACT

The logic of using summed radiocarbon ($^{14}$C) calibrations (cumulative probability density functions for large numbers of calibrated $^{14}$C dates) as proxies for past populations rests on the presumption of a proportional relationship between population size and the production, and subsequent preservation, recovery, and analysis, of $^{14}$C-datable material. Critiques of this approach have generally focused on the various problems that may undermine the validity of this assumption.

Here, instead, we presume a perfect correspondence between population size and the quantity of datable material produced at any given time, and explore the question of how well summed $^{14}$C calibrations can track demographic changes under such ideal circumstances. We introduce a method of generating a random sample of simulated $^{14}$C determinations, from a specified distribution, with variable data densities and measurement errors. In other words, we generate a random sample of $^{14}$C dates not from an ideal statistical distribution but rather using a defined population curve to determine the probability distribution from which the calendar dates of the simulated $^{14}$C samples are drawn. We generate simulated $^{14}$C ages for these samples, calibrate them, and sum those calibrations. We compare the resulting proxy population curve to the known population distribution from which it was generated, to see whether known population fluctuations are unambiguously visible on a proxy curve derived from $^{14}$C data sets that are realistic in terms of the number and precision of the $^{14}$C determinations included.

Results highlight 1) the critical role played by the magnitude and duration of any population fluctuation, and 2) the importance of sample size, and the reality that the numbers of samples required to detect significant population changes are generally far higher than those available to researchers proposing demographic reconstructions on the basis of literature searches for radiocarbon dates. We conclude that even if archaeological $^{14}$C data sets could be corrected for taphonomic filters and research biases, demographic signals would be difficult to distinguish from statistical noise in summed probability distributions. We suggest that simulation studies should be integral components of any attempt to reconstruct prehistoric demography from $^{14}$C dates.

## 1. Introduction

The last decade has seen a stream of publications, in peer-reviewed English-language archaeological journals, based on the premise that cumulative probability distributions of calibrated radiocarbon dates ("Sum distributions") are valid proxies for human populations (e.g., Gkiasta et al., 2003; Gamble et al., 2005; Shennan and Edinborough, 2007; Buchanan et al., 2008; Riede, 2009; Collard et al., 2010). The "Sum approach" has generally been driven by: the desirability of population time-series as an explanatory variable in analyses of cultural change (Shennan, 2009); poor chronological resolution of other potential population proxies; and especially the relative ease of data-mining for $^{14}$C results, compared to obtaining data on and interpreting more complex or qualitative variables, such as numbers of structures, quantities of artifacts, and settlement patterns. Even when a multi-proxy approach has been adopted, sums of calibrated $^{14}$C dates may appear to offer the most precise chronology for fluctuations in population inferred from other evidence.

In individual $^{14}$C calibrations, the height of a probability distribution at any calendar date corresponds to the probability that this

* Corresponding author. Tel.: +49 (0)431 880 5009.
*E-mail addresses:* dcontreras@ecology.uni-kiel.de, danielalexandercontreras@gmail.com (D.A. Contreras), jmeadows@leibniz.uni-kiel.de (J. Meadows).

http://dx.doi.org/10.1016/j.jas.2014.05.030
0305-4403/© 2014 Elsevier Ltd. All rights reserved.

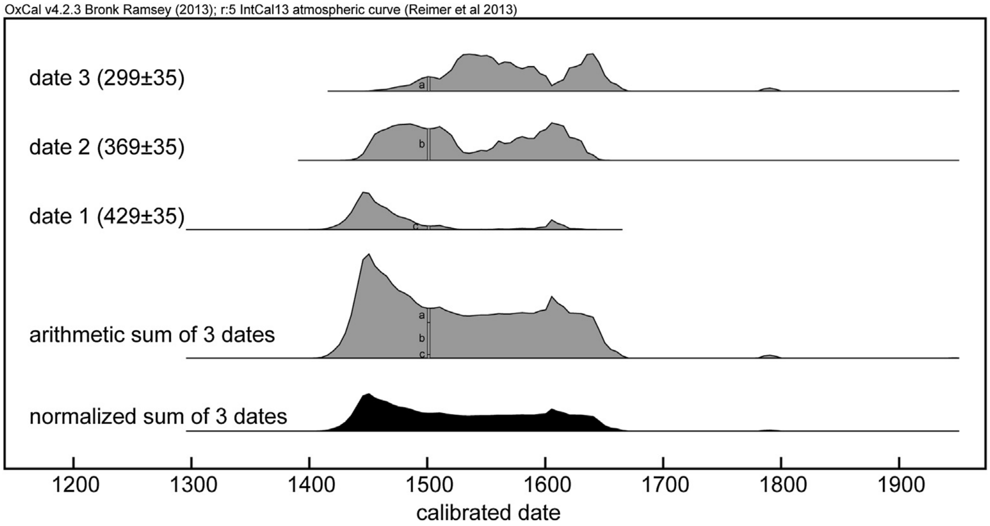

**Fig. 1.** A normalized Sum distribution (black), obtained in OxCal v4.2.3 (Bronk Ramsey, 2009a) by adding the heights of the calibrated distributions for dates 1–3 at every calendar date and normalizing the height of the arithmetic sum distribution to 1. A single arbitrary calendar date is marked to illustrate the summation of the three distinct probabilities.

is the true date of the material dated. Individual $^{14}$C calibrations can be combined by adding the height of each calibrated distribution at every calendar date, and normalizing the height of the resulting cumulative probability density function, which should contain the dates of all the samples concerned (Fig. 1). Sum distributions can easily be calculated using the calibration software CalPal (Weninger et al., 2007), Calib (Stuiver and Reimer, 1986–2014) or OxCal (Bronk Ramsey, 2009a).

In publications based on the "Sum approach", the peaks and troughs in a Sum distribution are regarded as proportional to the *number* of $^{14}$C samples of the corresponding calendar age. In effect, the Sum distribution is used as a proxy for the frequency distribution of dates; the real frequency distribution is unknowable, as the dates of individual samples are always ambiguous (Fig. 1).[1]

For even a real frequency distribution of dates to provide a good approximation of human population trends, we must assume that a) the dated samples are statistically representative of an underlying "population" of potential $^{14}$C samples, and b) there is a proportional relationship between past human population and the production of datable cultural material. Drawing on analogy with historic events as well as demographic and evolutionary theory, proponents of the Sum approach contend that population fluctuations are likely to be significant in (depending on the scale of investigation) the trajectory of a particular region or the human species, linked (as either cause or consequence) to, for instance, evolutionary bottlenecks, subsistence innovations, migrations, etc. As a result, they aim to identify population fluctuations of sufficient magnitude and duration to influence or explain important transitions in human history, identifying important (pre)historic events and enabling correlation with other exogenous events (e.g., climate changes, volcanic eruptions, asteroid impacts). Although some authors acknowledge that Sum distributions are at best imperfect proxies, others may give readers the impression that the major features of Sum distributions reflect population fluctuations quite precisely, in terms of both timing and scale. For example, Collard et al. (2010) and Kelly et al. (2013) use the slopes of spikes in a Sum distribution to calculate population growth rates, while Shennan et al. (2013:4) state that "In virtually all the regions examined here, there are significant demographic fluctuations and in most there are indications at certain points of population decline of the order of 30–60%".

The presumed link between the height of a *Sum* curve and the size of the corresponding human population has been criticized on methodological and theoretical grounds (e.g., Bayliss et al., 2007; Culleton, 2008; Chiverrell et al., 2011; Bamforth and Grund, 2012), as other factors than population size may create peaks and troughs in *Sum* curves (discussed in detail in Section 2). Nevertheless, Shennan (2013:305) maintains that "[the] key point is that even though a single date may have a broad calibrated range, the accumulation of the probability distributions of a large number of dates produces a high degree of chronological resolution making it possible to trace population fluctuations in considerable detail". The lure of prehistoric population proxies is such that publications which use the Sum approach continue to appear (Rieth et al., 2011; Hinz et al., 2012; Armit et al., 2013; Mulrooney, 2013).

To address the question of discerning real fluctuations in the detail of Sum distributions, we investigate here whether the Sum approach can identify the kinds of patterns which it aspires to find, in sets of simulated (artificial) $^{14}$C determinations that realistically represent the data density (average number of samples per year) available in archaeological $^{14}$C data sets. Specifically, we generate random samples of dates from distributions corresponding to specific human population scenarios (i.e., the probability of drawing a sample from a given date is exactly proportional to the population size in that year), simulate $^{14}$C determinations for those samples, sum the results, and compare those sums to the original population curves from which the simulated batches of dates were generated. The population scenarios that we explore are drawn from historical and archaeological data, and thus are realistic in their magnitudes and durations. We have selected population events that are notable not only for their scale, but also for their recognized historical consequences – i.e., exactly the sorts of events that Sum approaches hope to identify. It is clear from our results that even in rich data sets that lack any of the known sources of bias, the Sum

[1] Sum distributions can provide a reasonable approximation of changes in the temporal frequency of dated samples when a Bayesian model includes a high enough density of simulated data (Bronk Ramsey, 2001). In Bronk Ramsey's little-noticed example (2001: Fig. 3), the Sum distribution is the sum of posterior density estimates of the dates of simulated $^{14}$C samples, not of simple calibrations. The fact that phase boundaries are at exactly the right points (where the frequency of dates changes) helps to produce a Sum distribution that closely approximates the original distribution of sample dates, which are evenly spaced within each phase.

approach is not reliably effective at distinguishing population fluctuations of the scale interesting to archaeologists from statistical noise. We review below the requirements for such identification, before exploring simulations that assess the effectiveness of Sum approaches.

## 2. Logic of sum approaches: necessary assumptions and potential biasing factors

### 2.1. Sample production and survival

Any approach whose logic relies on diachronic comparison – in the case of Sum approaches to past population, of the quantities of datable material produced at different times – must confront the issue of taphonomy, the differential production and survival of potential $^{14}$C samples. “Correction” for destruction over time, using a simple exponential decay curve (i.e., assuming a constant rate of site destruction; Surovell and Brantingham, 2007; Peros et al., 2010), may provide a more realistic impression of the relative abundance of datable material in different periods, and may not even be necessary if the period studied is relatively brief compared to its absolute age. Moreover, comparisons within and between data sets representing different types of archaeological settings (e.g., settlements, cemeteries, monuments) have been suggested as a means of distinguishing between “real” and taphonomic patterns in the abundance of datable material (e.g., Riede, 2009; Collard et al., 2010; Hinz et al., 2012).

At a local level, taphonomy clearly produces spurious patterns; where bone collagen survives poorly, for example, an otherwise perfect $^{14}$C proxy might track the popularity of cremation burial, rather than the overall population. It is assumed, unrealistically in our view, that in large data sets, taphonomic filters will not discriminate between regions and site types, and that the quantity of surviving datable material should therefore reflect underlying trends in the production of datable material, trends which depend on the number of people alive to produce such material. More complex models, which attempt to account for palaeoenvironmental changes, such as sea-level rise, and for subsequent land-use, may provide more realistic indications of the survival of datable material from different periods, but again, there is a substantial body of literature based on the assumption that (other things being equal) taphonomic patterns will have a neutral effect on the resulting Sum distribution, rather than taking a more critical view (cf. Ballenger and Mabry, 2011, who present a case study in which other factors overwhelm production as a determinant of the abundance of datable material). Nevertheless, we presume for the sake of this study that an ideal case, in which taphonomy either is not a confounding factor or may be sufficiently accounted for, may exist.

### 2.2. Research intensity

At least as challenging is the issue of research intensity, already recognised by Rick (1987). Notwithstanding locally eclectic practices, the number of $^{14}$C samples dated in any region may best reflect that region's economic fortunes over the last 30 years, rather than its population in prehistory, but even within regions of comparable prosperity, perceptions as to the relative importance of different archaeological phenomena and the utility of $^{14}$C and other dating methods mean that resources will be unevenly directed towards dating different periods. A further complicating factor is that researchers collecting $^{14}$C results published in academic literature may be unaware of larger and perhaps less selective data sets generated by commercial archaeology (as Crombé and Robinson (2014) have recently observed). Attempts to “correct” Sum distributions for differential research intensity, either by summing the calibrated pooled means of $^{14}$C results from individual sites/site-phases (e.g. Shennan and Edinborough, 2007; Buchanan et al., 2008; Tallavaara et al., 2010) or by summing the calibrated dates for individual sites before summing the sums (e.g., Collard et al., 2010; Hinz et al., 2012) have never been justified on statistical grounds. They also clearly run counter to the assumption that larger populations would produce more datable material, as normalisation gives equal weight to every site or site-phase; Crombé and Robinson (2014) discuss how changing settlement patterns can then give rise to spurious fluctuations in Sum distributions.

### 2.3. Calibration and software issues

Another challenge faced by the Sum approach is to account for the effects of $^{14}$C calibration. Sum distributions of uncalibrated $^{14}$C results must, by definition, be misleading, as the relationship between calendar and $^{14}$C ages is not at all linear: if $^{14}$C samples are uniformly distributed in calendar age, there will be many more $^{14}$C results corresponding to calibration “plateaus” than to steeper sections of the $^{14}$C calibration curve (Reimer et al., 2013). Notwithstanding the suggestion by Hinz et al. (2012) that the $^{14}$C calibration curve does not produce false peaks and troughs in a Sum distribution of a sufficient number of calibrated $^{14}$C results, it can easily be demonstrated that simulated $^{14}$C determinations for samples whose calendar dates are evenly spaced (e.g. at 10-year intervals) will give a Sum distribution whose peaks and troughs directly correspond to calibration-curve wiggles (e.g., Chiverrell et al., 2011; Prates et al., 2013). The same pattern appears when the calendar ages of the simulated $^{14}$C determinations are randomly generated from a uniform distribution (e.g., Armit et al., 2013; Bleicher, 2013).

Some authors have subtracted an artificially generated “uniform” Sum distribution from the Sum “observed” in archaeological $^{14}$C data (e.g., Johnstone et al., 2006), or divided one distribution by the other (Hoffmann et al., 2008), in order to separate the calibration-derived peaks and troughs from the “residual” signal, but this can only work if the temporal frequency of archaeological $^{14}$C samples was also constant. Equally, subtracting a Sum distribution artificially generated by sampling an exponential distribution from a Sum distribution of archaeological $^{14}$C dates can only correct for calibration wiggles if the real dates of the archaeological $^{14}$C samples were exponentially distributed. More sophisticated computational approaches may eventually be successful in reducing the impact of calibration wiggles (e.g., Shennan et al., 2013; Kerr and McCormick, 2014), but certainly the problem must be addressed.

### 2.4. Sampling

Even if calibration, taphonomic and research intensity biases could be accounted for, however, data coverage remains a significant challenge. Any given $^{14}$C data set represents a single – and inevitably relatively small – sample of an unknown population of datable material produced during the period under investigation. The probability that this sample approximates the population from which it was drawn relates directly to sample size, and in the cases of spans of time more particularly to data density (i.e., number of $^{14}$C dates per given span of time). In order for robust patterns – which might be attributed to population changes – to be discerned in Sum distributions, data densities must be sufficiently high to reveal population changes of the posited magnitude and duration. The lack of a consensus definition for the magnitude and duration of either a “significant” population fluctuation or a “significant” wiggle in a Sum distribution remains a considerable conceptual hurdle, one to which we return below in evaluating the results of

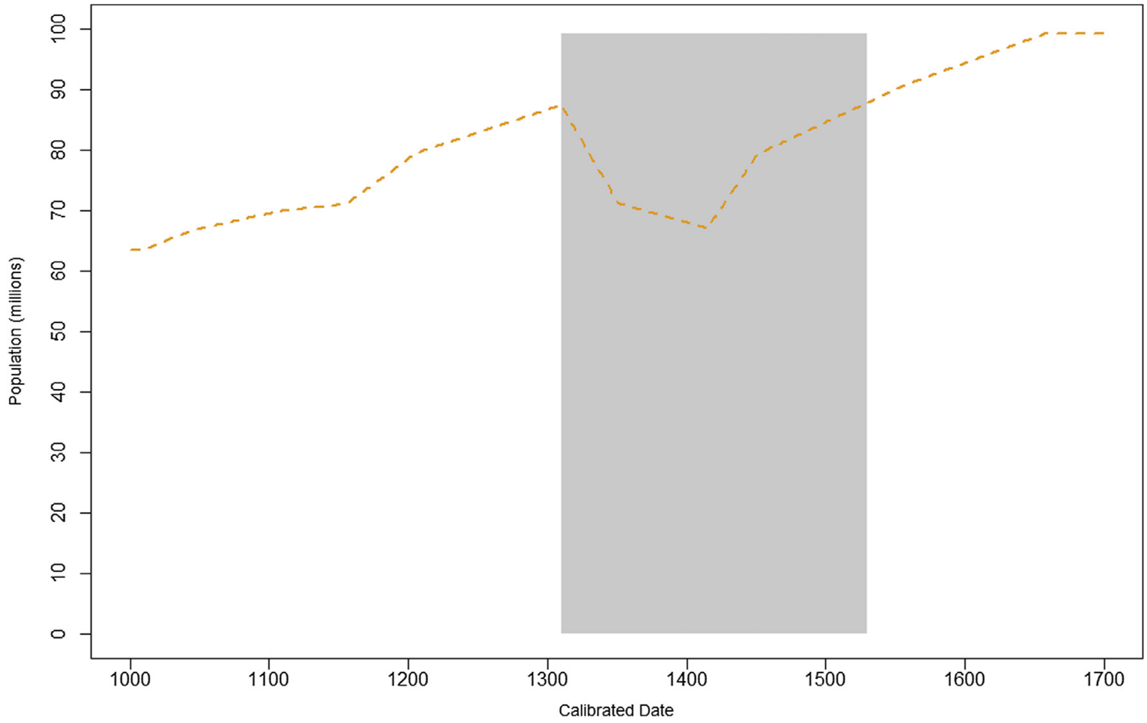

**Fig. 2.** European population curve (derived from Bennett's 1954 estimate as reported in Durand, 1977) that serves as the basis for simulations. Shaded area represents the population trough (i.e., the period between peak population in cal AD 1310 and recovery to that same level by approximately cal AD 1530).

our simulations. In general, while it may be possible to detect the simple case that populations tend to increase rapidly with the initial adoption of agriculture, or increase over time, prehistoric demography at this coarse level is already accessible through other proxies (i.e., site size and frequency). Detection of more complex, short-term, or subtle population shifts must overcome the problems of statistical scatter inherent in $^{14}$C data, even supposing that taphonomic and research intensity biases have been addressed. We explore the use of simulated data to address this problem below.

## 3. Artificial data

### *3.1. Simulation approaches to $^{14}$C data sets*

Whereas earlier publications (e.g., Gkiasta et al., 2003; Gamble et al., 2005) made no attempt to check whether “patterns” observed in Sum distributions might have arisen purely by chance, comparisons between Sums of archaeological and artificial $^{14}$C data are now published routinely. A common aim is to compare an archaeological Sum distribution to a uniform distribution (corresponding to a null hypothesis that a “population” does not change over time). As indicated above, this approach also has the perceived advantage of cancelling out spurious signals caused by $^{14}$C calibration. To simulate such a uniform distribution, Buchanan et al. (2011: 2120) calibrated 603 artificial $^{14}$C ages set at equal *$^{14}$C year* intervals between 13,000 and 8000 uncal BP, a method (repeated in Mulrooney, 2013) which, as well as disregarding the inevitable statistical scatter in $^{14}$C measurements, guarantees a much lower data density on calibration plateaus than during the brief steep sections of the calibration curve. A more realistic uniform Sum distribution can easily be implemented in OxCal, using simulated $^{14}$C ages (*R_Simulate*) for samples whose calendar dates are evenly spaced (e.g., Chiverrell et al., 2011). The *R_Simulate* function returns a different $^{14}$C age every time by randomly sampling the $^{14}$C ages that might be measured for the sample, given the calendar date and $^{14}$C measurement error set by the user.

To approximate an archaeological $^{14}$C data set, however, a simulation should account not just for $^{14}$C measurement scatter, but also for the effects of random sampling of *potential* $^{14}$C samples (of course, variability in taphonomy and research intensity mean that the archaeological $^{14}$C data do not represent an entirely random sample – but the premise of the Sum approach, as of many other approaches to archaeological data, is that it may be treated as such). In other words, rather than being spaced to match the null-hypothesis population time-series – typically, as described above, expressed as a uniform distribution – the calendar dates of simulated $^{14}$C samples themselves should comprise a *random sample* from such a distribution. The result will over-represent some years and under-represent others, reflecting the realities of random sampling and the effects of sample sizes. A Sum distribution of calibrated simulated $^{14}$C ages for samples whose calendar dates were randomly generated from a uniform distribution (Armit et al., 2013) therefore better approximates an archaeological data set. An exponential null-hypothesis curve may be interpreted as a presumption of steady population increase over time, but is typically intended to represent time-dependent survival of datable material generated by a constant population (Surovell and Brantingham, 2007). Arguably, there is no real justification for assuming a constant population as the null hypothesis, but operationally it is easier to implement and explain than a more complex pattern.

As well as calibration-related wiggles, therefore, in a realistic simulated Sum distribution there will inevitably be features (peaks or troughs) caused by random sampling and $^{14}$C measurement scatter, both of which are unavoidable in archaeological data. Any simulation exercise should be run repeatedly to see how variable its output is; features that do not appear in all simulations probably represent statistical noise. If a simulated *uniform* Sum distribution regularly contains features comparable to those in an *archaeological* Sum distribution with similar data density and measurement precision, we cannot attribute the features in the archaeological Sum distribution to population fluctuations (e.g., Bamforth and Grund, 2012). One way to assess the significance of any features seen in archaeological Sum distributions is to observe how often a simulated null-hypothesis Sum contains similar features (Shennan et al., 2013).

### 3.2. A flexible method for generating realistic artificial data

Although it is easy to generate random numbers from standard statistical distributions (e.g., uniform, exponential, Gaussian), it can be laborious to rigorously sample more complex distributions (e.g. Bayliss et al., 2013). We therefore constructed a flexible tool (`datesim`) in R (2013),[2] which generates a specified number of calendar dates by randomly sampling a user-specified custom distribution (corresponding for instance to a population curve with trends and fluctuations of particular amplitude and duration). The tool also generates individual measurement errors for the simulated samples, randomly sampled from a uniform distribution whose parameters can be set to reflect $^{14}$C measurement errors in the archaeological data set of interest. The R output is a .csv file which can be imported into OxCal, using either the *R_Simulate* function (which generates realistic $^{14}$C ages for samples, given their specified measurement errors, and calibrates the simulated $^{14}$C results), or *C_Simulate* (which generates a calendar date for each sample, based on its “real” date produced in R and its specified measurement error). Like *R_Simulate*, *C_Simulate* demonstrates the impact of measurement scatter, but it eliminates patterns imposed by the calibration curve itself. The Sum of the calendar/calibrated date distributions can then be compared to the population history from which it was derived.

### 3.3. Specification

Two steps are required to simulate dates from a particular population curve, or probability density function:

1) the approximation of the probabilities that will govern sample selection, which can be based on an empirical or artificial population curve (we used Plot Digitizer (http://plotdigitizer.sourceforge.net/) to convert published population curves to time-series data), and
2) the selection of a specified number of samples from a given calendar date range (corresponding to the range of the input curve, or subset thereof) and the corresponding randomly-generated measurement errors (drawn from a uniform distribution with a specified maximum and minimum; it would be possible to specify other distributions if desired). The annotated R code is provided below.

*Step 1*[3]

```
#Digitize the population curve so that it is represented as a series of
values (pop$V1) corresponding to dates (pop$V2) (a table of population
values will already have this form), then interpolate the missing values
using approx(). The frequencies of values in the resulting data frame (here
“popinterp”) will serve as the probabilities (here “probs”) for sample() in
Step 2.
popinterp<-approx(x=pop$V1, y=pop$V2, method="linear", xout=1000:1700,
rule=2)
probs<-popinterp$y/sum(popinterp$y)
```

*Step 2*

```
#Create a ‘datesim’ function that takes the inputs described below and
draws a random sample with replacement using the probabilities provided by
the interpolated population curve calculated in Step 1 (here “probs”).  The
output from ‘datesim’ is a matrix with three columns of data corresponding
to an arbitrary sample designation (here “A1”, “A2”, etc), the randomly
generated date, and a randomly generated measurement error drawn from a
uniform distribution whose parameters may be specified (in this example,
between 20 and 40; other distributions may also be specified).  The results
may be exported via write.csv() and then copied directly into OxCal using
the Import tool.
#num = length(sample)
#oldest = min(sample)
#youngest = max(sample)
#format can be either BP or BC/AD
datesim<-function(num, oldest, youngest)   {cbind(paste("A",1:num, sep=""),
sort(sample(c(oldest:youngest), num, replace=T, prob=probs)),
round((runif(num,20,40)),0))}
```

[2] The application described here also uses the ‘`stats`’ package (“The R Stats Package” – http://stat.ethz.ch/R-manual/R-patched/library/stats/html/00Index.html).

[3] The type of interpolation is definable by ‘`method`’ in `approx()`; “`approx` returns a list with components `x` and `y`, containing `n` coordinates which interpolate the given data points according to the `method` (and `rule`) desired.” (http://stat.ethz.ch/R-manual/R-patched/library/stats/html/approxfun.html). In the case that rather than a series of data points a mathematically defined curve were available, Step 1 could be skipped and the curve could govern the sampling probabilities directly (defined as ‘`prob`’ in `sample()`).

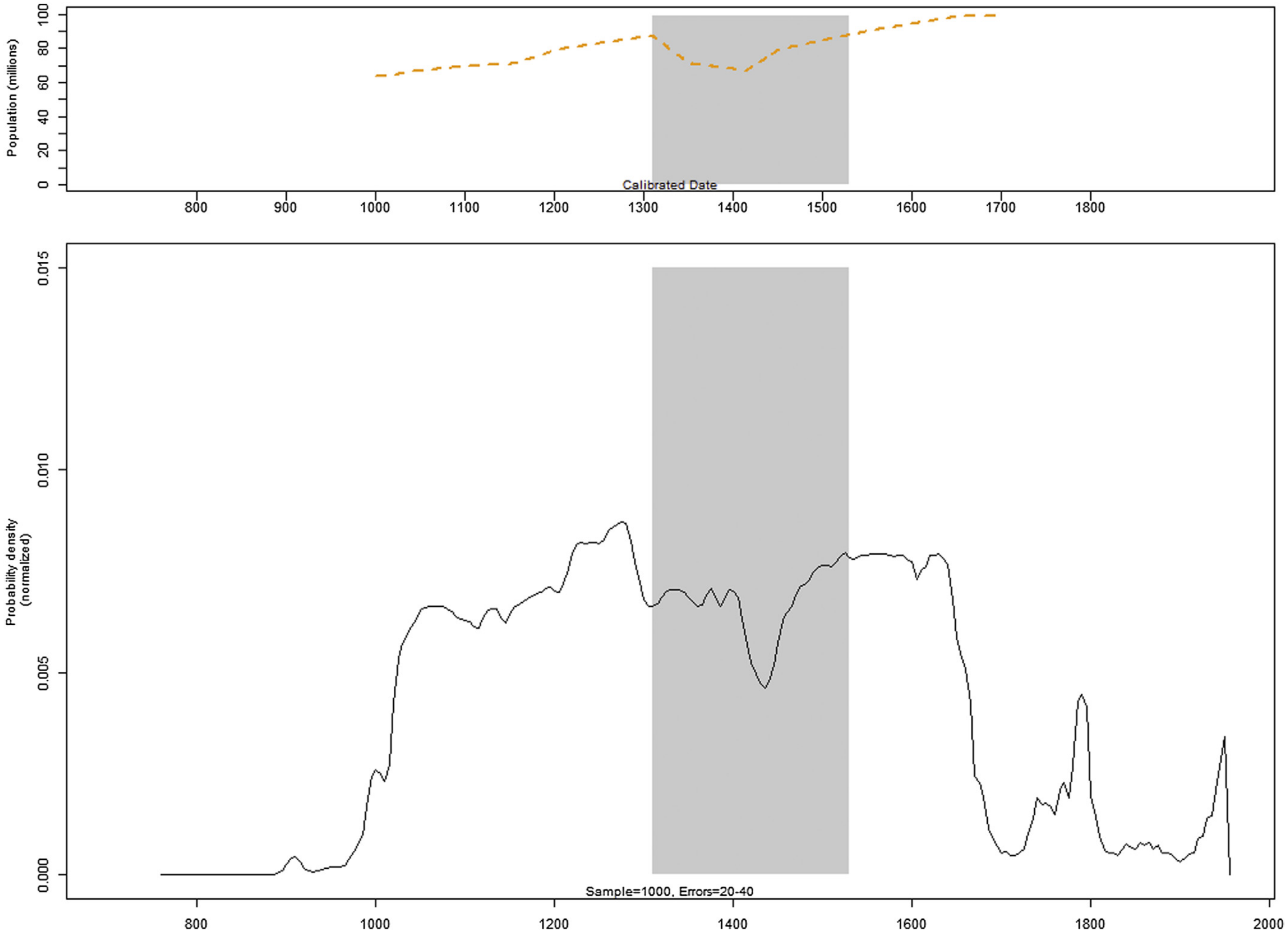

**Fig. 3.** Randomly selected (from the five sets of samples generated) plot of 1000 samples, run through *R_Simulate*.

### 3.4. *Worked examples*

We describe two the examples here, each deriving simulated dates from extant population data corresponding to a period of recognized population change and historical impact – i.e., the sorts of periods and population changes which Sum approaches aspire to describe.

#### 3.4.1. *"Bring Out Your Dead": demography and significance of the Black Death in 14th century Europe*

The Black Death was a series of epidemics of the bubonic plague that devastated Eurasia in the 14th century AD, and is particularly well-known from European historical sources, allowing realistic reconstructions of its cumulative impact on population. Moreover, the population reductions resulting from the Black Death had a long-term economic impact which can be traced in succeeding centuries (cf. Pamuk, 2007; among many), and it is manifest in the archaeological record, e.g. in mass burials of plague victims (Antoine, 2008) and in the abandonment of thousands of villages and farms across Europe (Yeloff and van Geel, 2007 Antonson, 2009). The shortage of agricultural labourers led to reforestation on such a scale that its impact may even have caused the notable decline in atmospheric $CO_2$ concentration in the later 14th century (Ruddiman, 2003). Given such a population impact and such important repercussions, it represents the type of demographic event we would hope to be able to detect unambiguously in a Sum distribution.

For the sake of this exercise, we used Bennett's (1954; summarized in Durand, 1977) population estimates to create a European population curve spanning the years AD 1000–1700. Although estimates of European population in this period vary, and are more accurate and precise for some countries than others (cf. Livi-Bacci, 1999; Pamuk, 2007), for our purposes the magnitude and duration of the population crash are important rather than the details. Our simulation is concerned not with *accurately* modelling this particular demographic event, but rather with attempting to reproduce the selected population curve using the Sum approach.

In our population curve, after rising relatively steadily for the first three centuries of this period, population declined abruptly between AD 1310 (87 million) and 1350 (71 million), and further declined to 67 million by AD 1415, before recovering to 79 million (AD 1451) and finally overtaking its pre-Black Death peak in c. AD 1550. Thus the population trough was both deep and prolonged (see Fig. 2); the magnitude and duration were large enough to have significant historical effects and the event thus provides a realistic analogue for the kinds of population fluctuations in prehistory in which we might be interested.

*3.4.1.1. Simulation parameters.* Five batches of samples of 200, 1000, and 2000 dates were drawn from this distribution, with

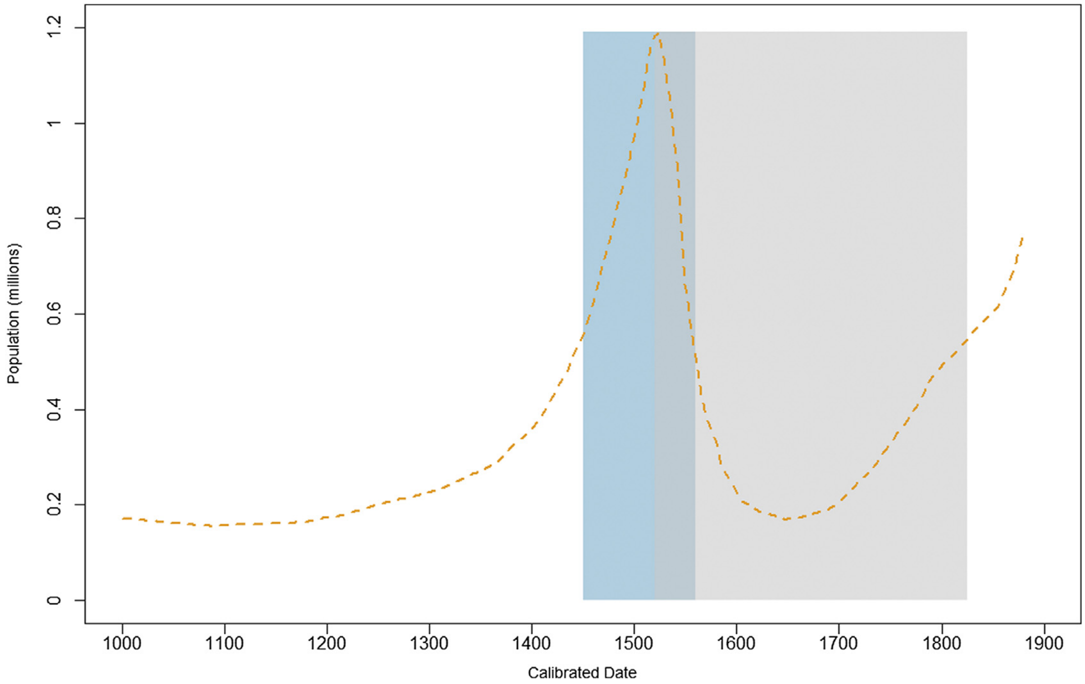

**Fig. 4.** Population of the Basin of Mexico, derived by averaging McCaa's (1995) estimates. The blue rectangle demarcates the population peak between cal AD 1450 and 1560 (by which time population had dropped to 1450 levels – an arbitrary reference point – again) and the grey rectangle the post-contact trough, beginning in cal AD 1520 and continuing until 1825, when population had recovered to 1450 levels. These rectangles delineate two distinct population phenomena: a discrete and ephemeral spike, and a crash followed by partial recovery. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

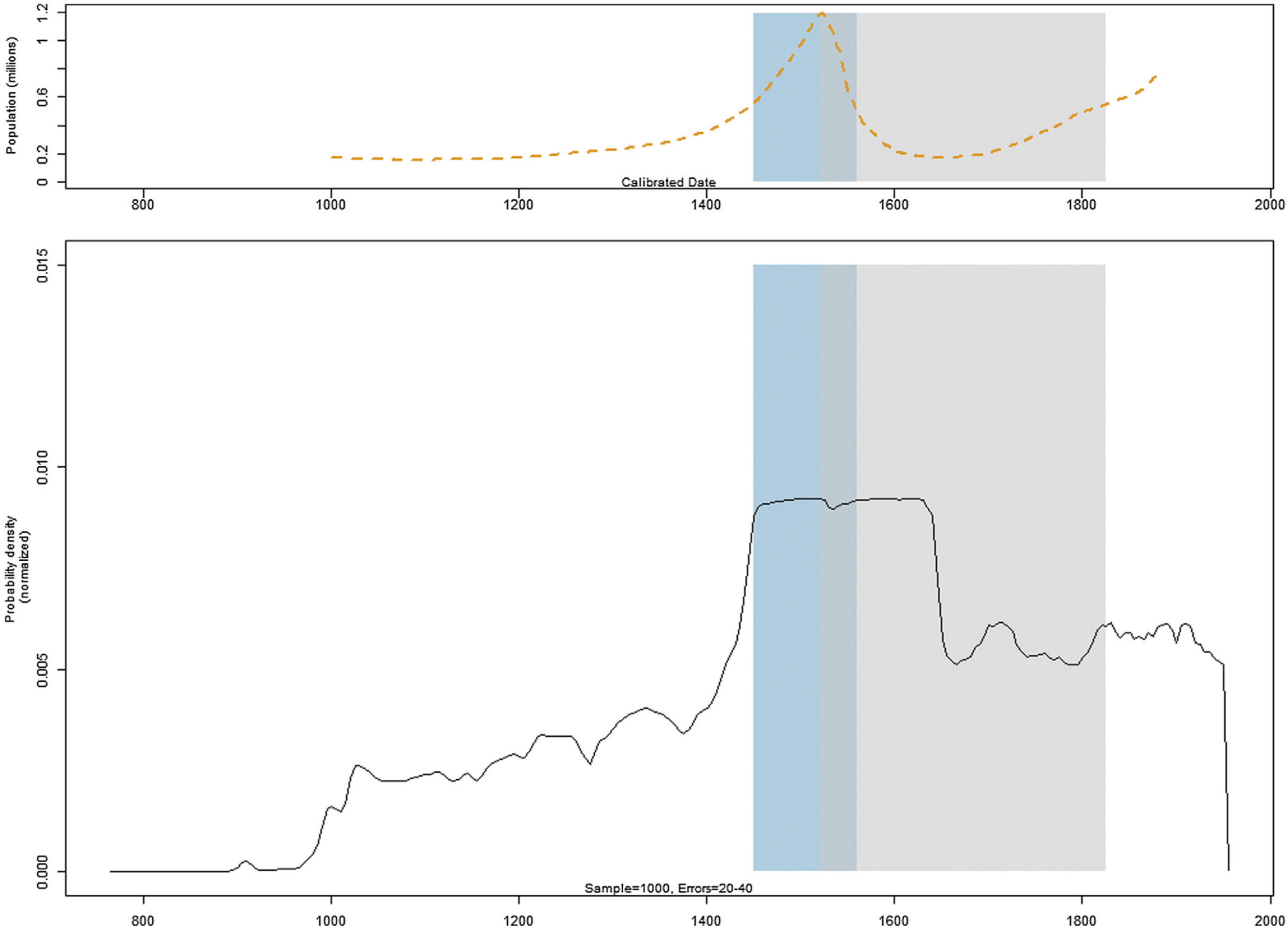

**Fig. 5.** Randomly selected (from the five sets of samples generated) sum of 1000 samples, generated through *R_Simulate*.

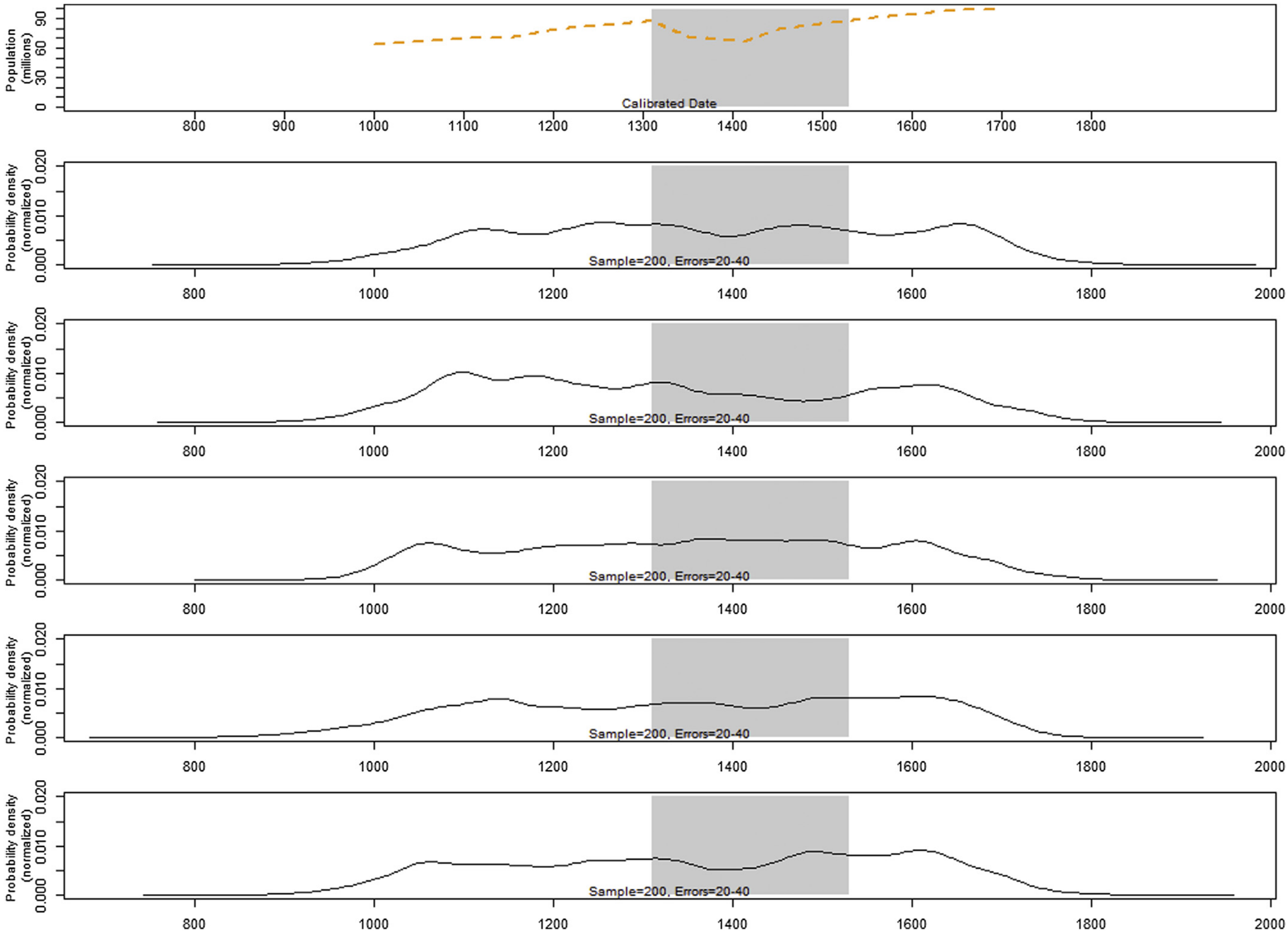

**Fig. 6.** Effects of random sampling: five different random samples of 200 dates from the cal AD 1000–1700 distribution shown in the top panel; we have used *C_Simulate* to reproduce realistic measurement scatter without introducing additional patterns derived from radiocarbon calibration.

**Table 1**
Comparison of $^{14}$C data density in a selection of typical publications that use fluctuations in Sum distributions as proxies for fluctuations in demographics or activity (we have excluded studies that focus primarily on long-term trajectories).

| Study | Number of dates[a] | Approximate period[b] (calendar years) | Approximate data density (dates/year) [Effective density[c]] |
|---|---|---|---|
| Armit et al., 2013 | 1554 | 1200 cal BC–cal AD 400 (1600) | 1.0 |
| Buchanan et al., 2008 | 1509 (628 dates or pooled means) | 15–9 ka cal BP (6000) | 0.25 [0.1] |
| Collard et al., 2010 | 4246/2129 (sum of 1762 sums) | 8000–4000 cal BC (4000) | 1.1/0.5 [0.4] |
| Gamble et al., 2005 | 2255, split into subsets of 28–500 dates | "25–11 ka calibrated radiocarbon years" (14,000) | 0.15 [0.002–0.04] |
| Gkiasta et al., 2003 | 2600 (subsets often <100) | 10–5 ka cal BP (5000) | 0.5 [0.02–0.2] |
| Hinz et al., 2012 | 3176 (subsets of 22–576 dates, summed by site) | 4500–2500 cal BC (2000) | 1.6 [0.01–0.1] |
| Mulrooney 2013 | 313 (subsets <200) | cal AD 1200–1800 (600) | 0.5 [0.1–0.3] |
| Riede, 2008 | 139 (subsets of 2–53 dates) | 15–11 ka cal BP (4000) | 0.03 [<0.01] |
| Rieth et al., 2011 | 303 | cal AD 1200–1600 (400) | 0.75 |
| Shennan and Edinborough, 2007 | 2311 (compare sets of 996, 213, and 366 dates; compare 294, 66, and 162 pooled means) | 7000–2000 cal BC (5000) | 0.4 [0.01–0.2] |
| Shennan et al., 2013 | 13,658 (compare sets of 281–1732 dates, summed within 151–928 'bins') | 10–4 ka cal BP (6000) | 2.3 [0.04–0.23] |
| Tallavaara et al., 2010 | 1789 dates (1160 pooled means; compare subsets of 238–513) | 11–1 ka cal BP (10,000) | 0.2 [0.02–0.05] |
| This study | 200, 1000, 2000 simulated | cal AD 1000–1700 and 1000–1800 (700–800) | 0.29, 1.43, 2.86 and 0.25, 1.25, 2.50 |

[a] Criteria for excluding misleading dates vary between studies, but (as Shennan and Edinborough, 2007 point out) it is impossible to eliminate all misleading dates in archaeological data sets. Several studies combine some dates before summing, either by calculating a pooled mean of dates from one site-phase, or summing the dates from one site-phase and treating their sum as a single date. Most studies also split their data sets to compare Sums of subsets of the data.

[b] Exact timespan is often not specified, and criteria for including marginal dates will vary between studies.

[c] Approximate data density in Sums of smaller subsets of dates compared in the study, averaged over the relevant date ranges (which may be shorter than the overall period of interest).

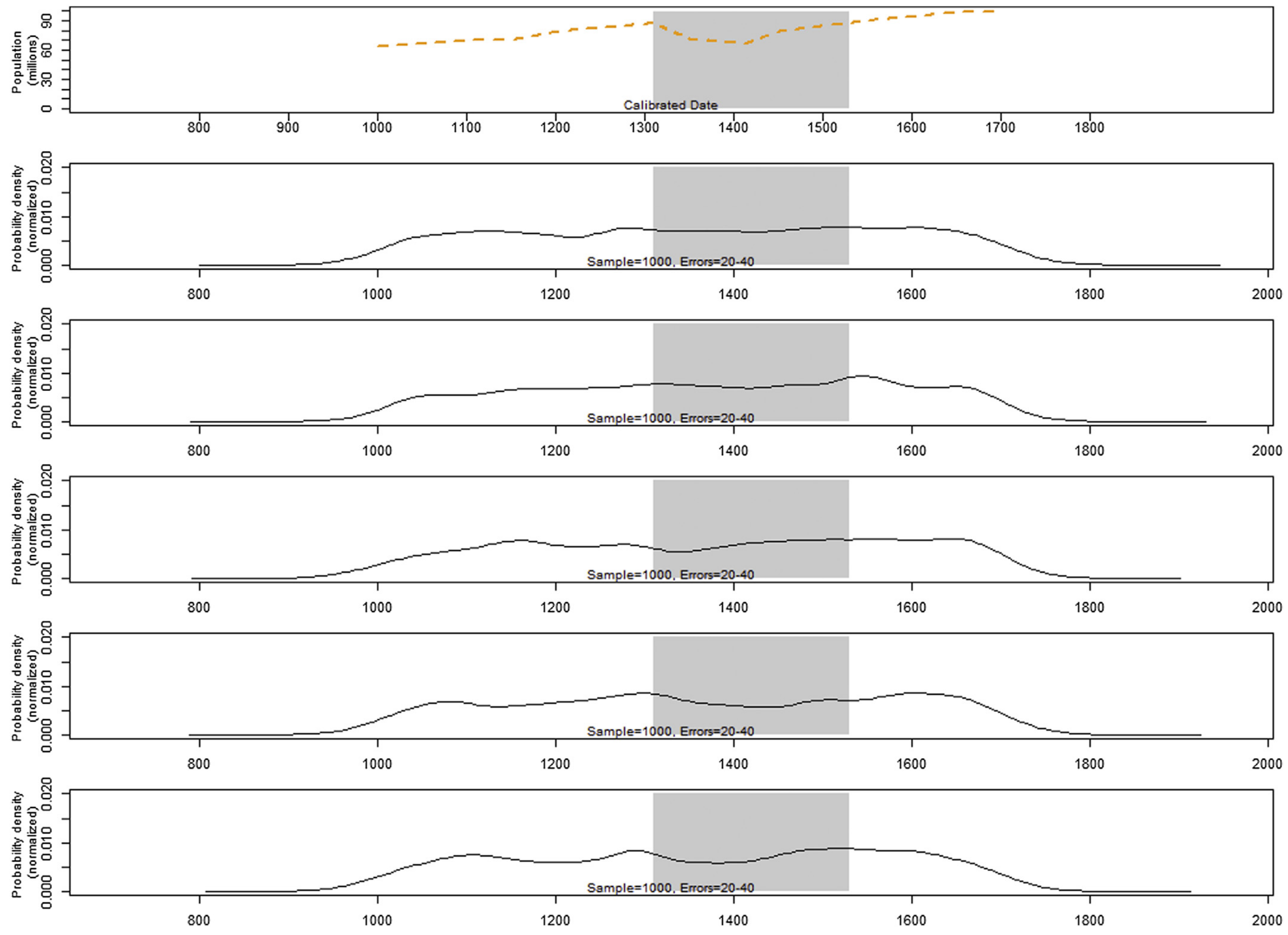

**Fig. 7.** As Fig. 6, but with sample sizes of 1000 rather than 200; even at this sample density (on average, 1–2 samples per year), the trough corresponding to the Black Death is barely visible in the date sum. As can be expected, variability between samples decreases as sample size increases (compare with Fig. 6).

measurement errors uniformly distributed between 20 and 40 years (a further selection of the same samples but with errors between 20 and 100 years was also generated to examine the effects of measurement error, discussed below), with samples drawn from the period AD 1000–1700. Both *C_Simulate* and *R_Simulate* functions were used to generate the sums that are discussed below.

*3.4.1.2. Results.* We illustrate the results of this simulation with one of the randomly-generated batches of 1000 samples (Fig. 3), already optimistic in both its measurement precision (i.e., measurement errors of 20–40 years), and in its data density (>1 date/year on average). What might we conclude on the basis of this assemblage of $^{14}$C dates? The dip in the sum curve at approximately cal AD 1450 might lead one to suppose that a decline during the period of the Black Death is detectable. A gradual increase in the centuries leading up to the plague period is also detectable, but after cal AD 1450 the “population” never seems to reach pre-plague levels again, much less surpass them. In addition, the plague trough seems to contain two distinct crash episodes, one at about cal AD 1300 and the other at about cal AD 1420. The calendar range of the Sum is also of note: even in the absence of samples introducing spurious ages, the combination of the calibration curve and $^{14}$C measurement uncertainty produces a sum that is not confined to the cal AD 1000–1700 range from which the original samples are drawn. Clearly these tails are to be disregarded, but distinguishing tails from the range of the sampled population presents an additional hurdle in the case of unknown populations.

This example, illustrating a realistic (and optimistic, in that errors are small and data density high) case, highlights a problem to which we will return below: how is the aspiring demographer to determine which fluctuations to recognize as related to population? The trough at approximately cal AD 1450 does indeed appear to register in some way the effect of the Black Death, but without prior knowledge would we recognize it as such? As the height of the sum curve does not have a direct relationship to any *particular* cultural or demographic variable, assessing the implications and scales of visible fluctuations remains a subjective exercise (a point to which we return in Section 4.5). We will also return below, in the Section 4, to the effects of varying the data density and magnitude of measurement errors.

### 3.4.2. *“…now a very great and notable fraction of the people are gone”[4]: 16th century demographic catastrophe in the Americas*

One of the major structuring factors of the modern world was the catastrophic impact of European diseases on the indigenous inhabitants of the Western Hemisphere. With population losses

[4] Ceynos 1565, quoted in McCaa 1995:430.

estimated at up to 90% in the 16th century (cf. Livi-Bacci, 2006; Lovell, 1992; McCaa, 1995, among many), and a long, slow demographic recovery, Central Mexico from AD 1200–1700 provides a historically attested example of the kind of extreme population fluctuation that Sum approaches seek to identify in prehistory. Using the mean of McCaa's (1995) high and low population estimates for the Basin of Mexico, we complement the example of the Black Death in Europe with this event of greater magnitude and duration. In addition to the massive crash associated with the impact of European diseases, this period in the Basin of Mexico is also notable for the rapid population growth preceding European arrival, the result of regional population aggregation associated with the political ascendancy of the Triple Alliance (see Fig. 4).

*3.4.2.1. Simulation parameters.* Five batches of samples of 200, 1000, and 2000 dates were drawn from this distribution, with measurement errors uniformly distributed between 20 and 40 years (a further selection of the same samples but with errors between 20 and 100 years was also generated to examine the effects of measurement error, discussed below), with samples drawn from the period AD 1000–1800. Both *C_Simulate* and *R_Simulate* functions were used to generate the sums that are discussed below.

*3.4.2.2. Results.* The discrete spike in population in this data set (Fig. 4) represents an ideal case for detection by *any* demographic proxy (indeed, the timing and magnitude of the 14th–15th century population increase are inferred from archaeological settlement survey (cf. Parsons, 1974)). The sharp 15th century rise reflects the florescence of the Triple Alliance, while the subsequent disastrous crash resulted from the arrival of the Spanish in the Basin of Mexico. It should thus perhaps come as no surprise that this spike is also evident in the summed simulated $^{14}$C data, with both a sudden increase and a sudden decrease apparent; we illustrate this with a randomly selected batch of 1000 samples in Fig. 5. As noted previously, however, there is no clear methodology for interpreting the Sum distribution. Although the general pattern is visible, various problems are also apparent: the 18th-century recovery is not evident, the timing, duration, and magnitude of the 15th-century spike are inconsistent with those in the original population curve, and the pre-1400 population curve is characterized by spurious variability. We address these issues in Section 4.

## 4. Discussion

Several parameters can influence the shape of the summed $^{14}$C curves; one of the advantages of working with simulated data is the ability to explore the effects of varying these parameters individually. Here we examine the effects of random sampling, data density, measurement uncertainty, and the calibration curve with reference to various data sets generated using our two artificial

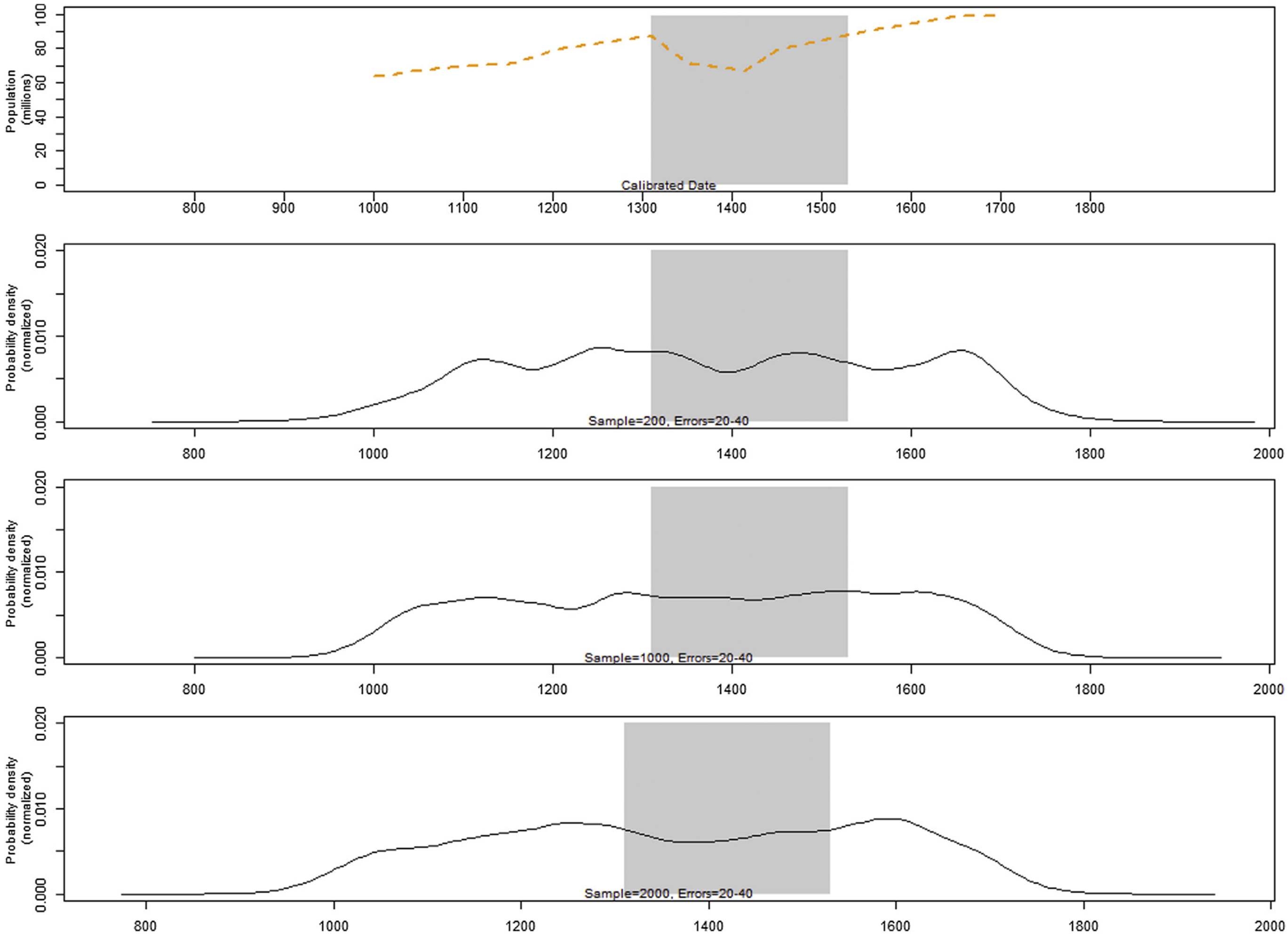

**Fig. 8.** European population cal AD 1000–1700, with sums derived from samples of 200, 1000, and 2000 dates, all with errors between 20 and 40. Sums are based on *C_Simulate* so that the variation between curves is primarily caused by sampling.

scenarios, before turning to an evaluation of some of the problems that may hamper interpretation of these Sum distributions as population proxies.

### *4.1. Random sampling*

The effects of random sampling are evident in Fig. 6. Even without the meaningless peaks and troughs in Sum distributions which stem from the calibration of $^{14}$C measurements (i.e., using OxCal's *C_Simulate* function to generate realistic measurement scatter, instead of *R_Simulate*), with 200 samples over 700 years (a data density better than many published examples; see Table 1) the effects of sampling are striking. Not only is the departure of these curves from the population distribution from which they are derived evident; the variability between samples is also notable: the most prominent fluctuations in each curve are not visible in most of the others. A Sum distribution of archaeological $^{14}$C data would correspond to only *one* of these curves, and although subsamples might be compared to one another, the underlying population distribution remains fundamentally unknown. As a result, it is difficult if not impossible to determine which, if any, of these fluctuations represent real demographic changes. Reduced intersample variability is evident as data density increases (Fig. 7), but it remains the case that sampling can have a notable effect on the structure of the Sum curves, and the vagaries of sampling alone may be sufficient to introduce or obscure patterns that may be taken as demographic indicators. This is the case in both practical and theoretical terms.

In practical terms, it introduces the problem that even in an ideal world, patterns in $^{14}$C sums may result from factors other than population fluctuations. As we discuss below, random sampling can produce both false negatives (failures to detect real population fluctuations) and false positives (features in Sum distributions that may be mistaken for results of demographic changes but in fact relate to other factors).

In theoretical terms, as any set of $^{14}$C dates comprises a single sample – of many theoretically possible – sampling introduces a degree of uncertainty into any assertion of the relationship of a given set of $^{14}$C data to the population of datable material from which it was drawn. The populations (of datable material) from which these samples are drawn remain fundamentally unknown, and some uncertainty is therefore inherent in any linkage of a $^{14}$C data set to the prehistoric population of datable material originally produced (see Drennan, 2009:93–95 on the relationship between archaeological samples and target populations). This uncertainty is independent of any putative relationship between production of datable material and human demography.

### *4.2. Data density*

Larger samples more closely approach the original population from which they are drawn – a fundamental tenet of statistics.

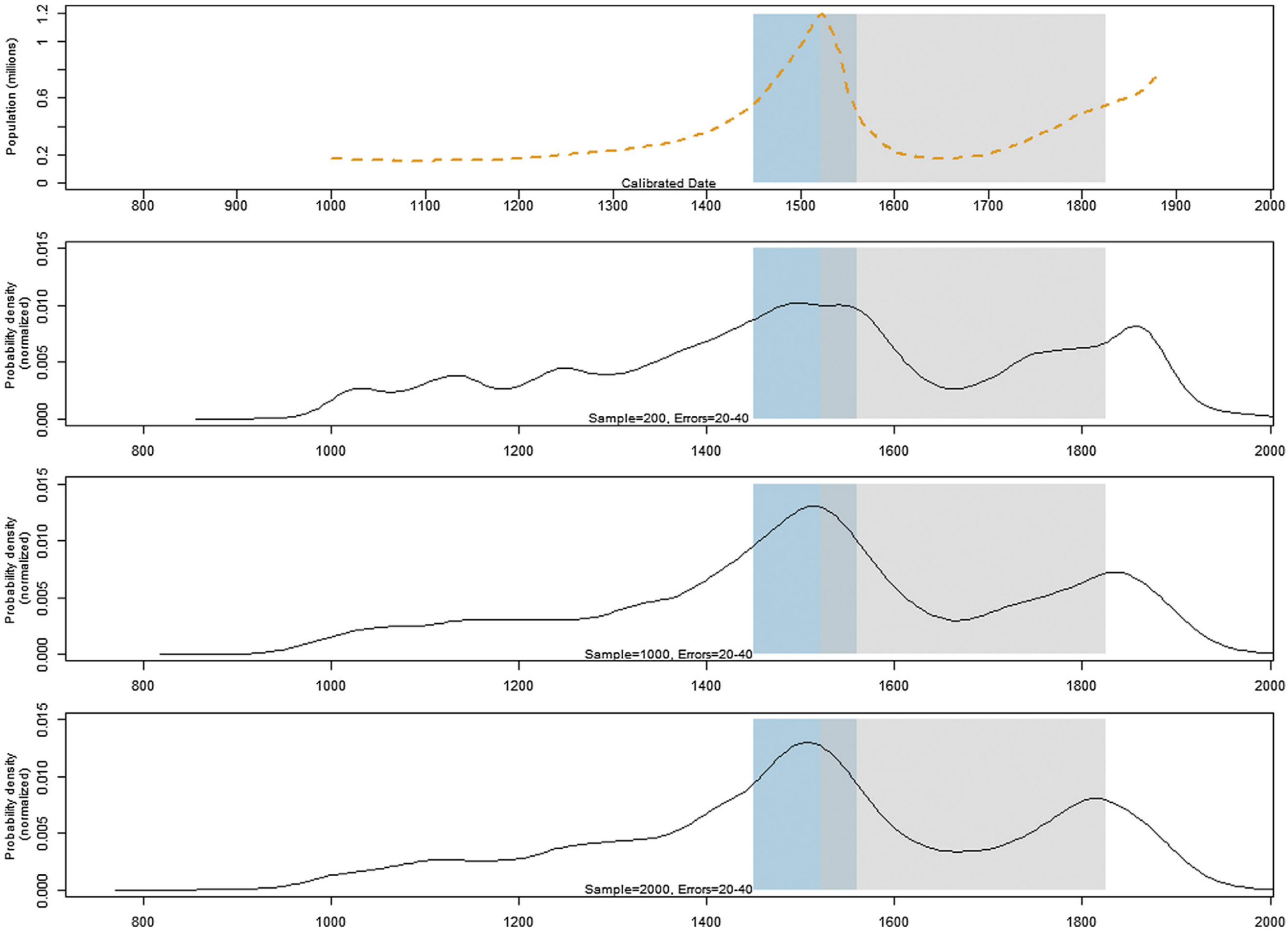

**Fig. 9.** Basin of Mexico population cal AD 1000–1800, with sums derived from samples of 200, 1000, and 2000 dates, all with errors between 20 and 40. Sums are based on *C_Simulate* so that the variation between curves is primarily caused by sampling.

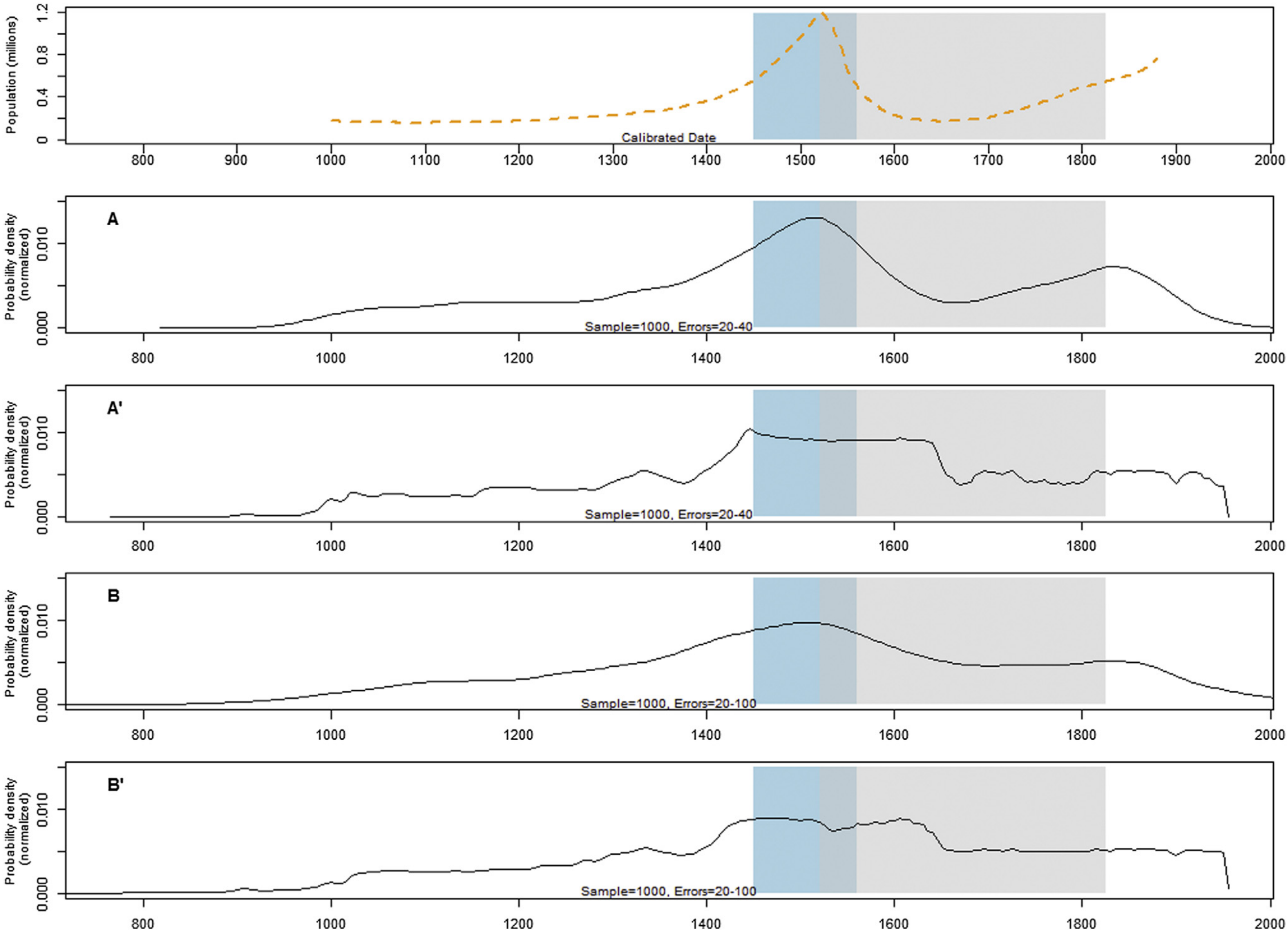

**Fig. 10.** Basin of Mexico simulations; these show the same data with varying errors (20–40 in A and A′; 20–100 in B and B′) and with *C_Simulate* (A and B) and *R_Simulate* (A′ and B′).

However, prescriptive approaches to necessary sample size (e.g., Hinz et al., 2012; Williams, 2012; cited by, e.g., Whitehouse et al., 2013) must take into account the span of time and the magnitude and duration of fluctuations or trends which the study seeks to detect. Assertions about necessary and/or sufficient numbers of $^{14}$C dates should be assertions about data density, rather than numbers of samples.

We illustrate the effects of data density with examples from both of our simulated cases (see Figs. 8 and 9). In the absence of other confounding factors (i.e. $^{14}$C calibration), increasing sample size – i.e., data density – clearly improves the fit of sum curves to the population curve. However, even the highest data density which we have tested – 2000 dates over a span of 700 years, or 2–3 dates/year – the Black Death is far from obvious. Our tested data densities range from an average of 0.25/year (i.e., one date per four years) to 2.86/year; most published examples use archaeological data sets with 0.5–1.5 dates/year, although the data are then often pooled or divided geographically, giving real average data densities of 0.1 dates/year, or even less (see Table 1).

### 4.3. Measurement precision

The precision of the $^{14}$C determinations that serve as the basis for the Sum is also a vital parameter. The general basis for our simulations here is a realistic best-case scenario of errors ranging from 20 to 40 years, which is typical of recent $^{14}$C determinations on relatively young samples (mid-Holocene or later). If – as is likely in incorporating dates from older data sets – errors are instead simulated as ranging from 20 to 100 years, the resulting Sum distributions are smoother and fluctuations less likely to be detected (Fig. 10; also see Culleton, 2008). The inclusion of legacy data – a feature of all the published examples cited above – thus increases data density but blurs any features.[5]

A further complication with regard to linking $^{14}$C sums to past populations is the likelihood of some mismatch between sample age and the date of the target event. Problems such as wood-age offsets, residence time, etc. reduce the effectiveness of measurement precision, but not symmetrically (see discussion in Bronk Ramsey, 2009b). Even when data sets are subjected to rigorous chronometric hygiene (Spriggs, 1989), we are likely often overestimating their precision – again a problem endemic to the use of legacy data. This can contribute to blurring of signals (e.g., by producing a spread of dates for a single event, giving the appearance of longer duration and/or making a signal even harder to detect in a Sum of $^{14}$C dates) or attribution of a precise but inaccurate date for a particular event (i.e., in the case that an event is represented only by $^{14}$C dates which do not accurately match the event). Inasmuch as the correspondence between particular events

[5] Smoothing algorithms (e.g., rolling means) have also been applied to Sum distributions in some studies (e.g. Hinz et al., 2012; Kelly et al., 2013), though there is a clear risk of effacing real signals as well as noise and in spite of the absence of any particular statistical justification. We have not investigated what effect they might have on our simulated Sums.

and dated material can be fundamental to the interpretation of Sum approaches (if, for instance, they attempt to relate a population fluctuation to an exogenous event; e.g., Buchanan et al., 2008; Riede, 2009), erosion of precision by any of these factors presents an obstacle that must be negotiated.

### 4.4. *Calibrated dates*

Using *R_Simulate* rather than *C_Simulate* – i.e., building Sum distributions from irregular calibrated probability density functions rather than Gaussian ones which only take into account measurement error (see Section 3.2, above) – logically produces spikier, less regular Sum distributions. We illustrate these effects for both simulations (Fig. 11).

Unless a researcher can completely compensate for such irregularities, which (we believe) is impossible without knowing the underlying distribution of the dates of the samples (the very pattern a researcher hopes to reconstruct from the Sum distribution), calibration creates a further challenge: the distortions of the Sum distribution caused by calibration will depend on the calibrated date range under consideration. This is illustrated in Fig. 12, in which the same Basin of Mexico simulated Sum distribution (Fig. 12A) has been recreated 2000 (Fig. 12B) and 5000 (Fig. 12C) years earlier, in the mid-1st and mid-4th millennia cal BC – periods chosen for comparison with recent publications, by e.g. Armit et al. (2013) and Hinz et al. (2012). In the first case, the calibration plateau between 750 and 400 cal BC spreads the peak region of the Sum distribution earlier and later; in the second, there is a peak at the right date, but a spurious peak as high as the "true" peak appears more than a century later, and is followed by a steeper decline.

### 4.5. *Resulting interpretive problems: recognizing patterns, and separating patterns from noise*

The interplay of these parameters can introduce several problems, even if we ignore the various factors that can attenuate or complicate the relationship between prehistoric population and production/preservation/analysis of datable material. To illustrate these problems, we focus on two simulations of 1000 samples, one each from the European and Mexican scenarios (Figs. 13 and 14 respectively). Using *R_Simulate*, we intentionally generate higher data densities than available to most researchers, and use optimistically precise measurement errors of 20–40 years for each date. Nevertheless, several pitfalls are evident. We do not to suggest that the problems discussed below are inevitable – but they *may* occur, partly as the result of sampling bias, even at high data densities. Their potential occurrence, in the case of inevitably singular and not (easily, if at all) replicable archaeological samples,

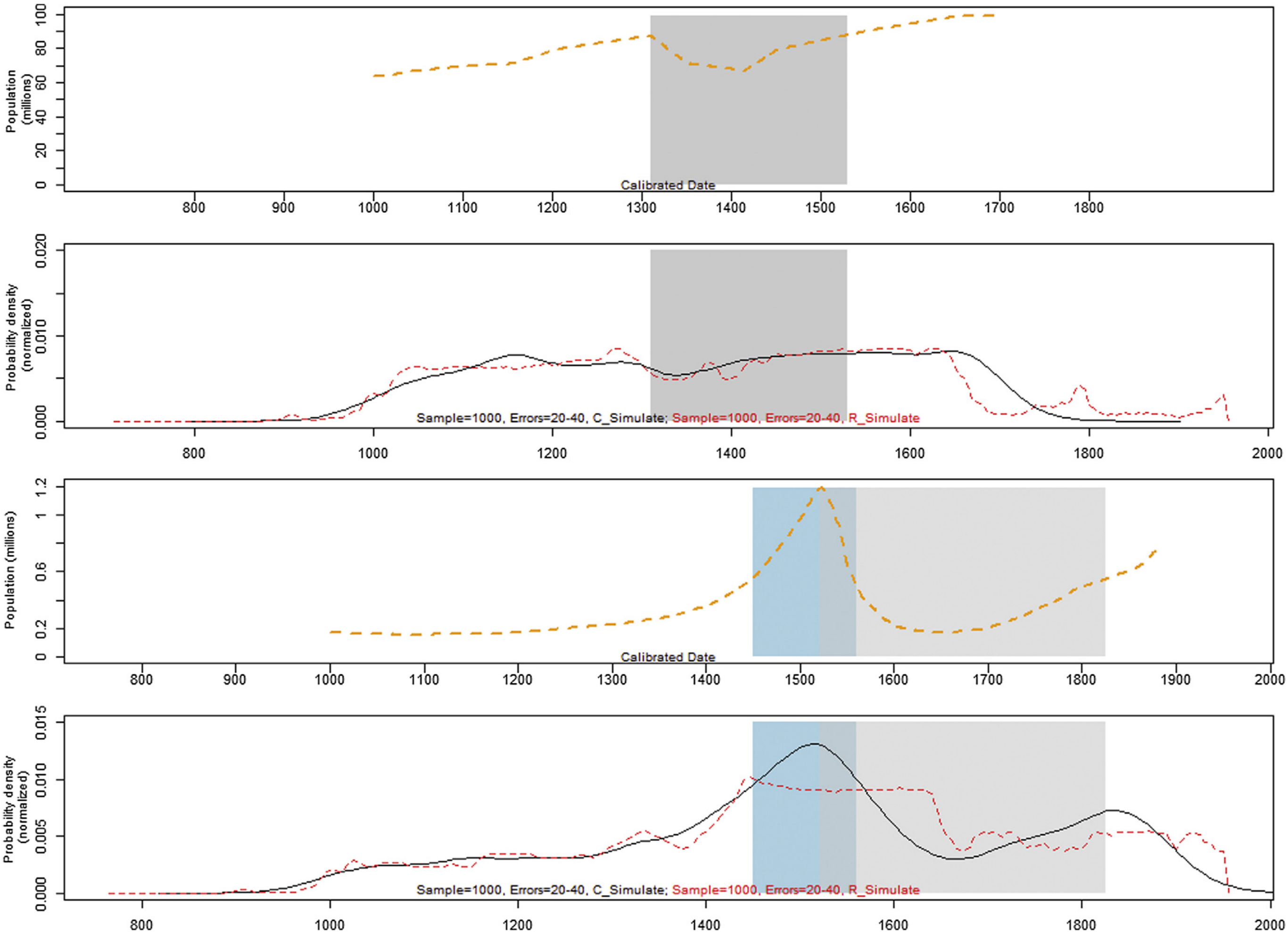

**Fig. 11.** Effects of the calibration curve on the Black Death and Basin of Mexico simulations. In each case the two lines are derived from the same sample of 1000 dates, run with *C_Simulate* (solid black line) and *R_Simulate* (dotted red line). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

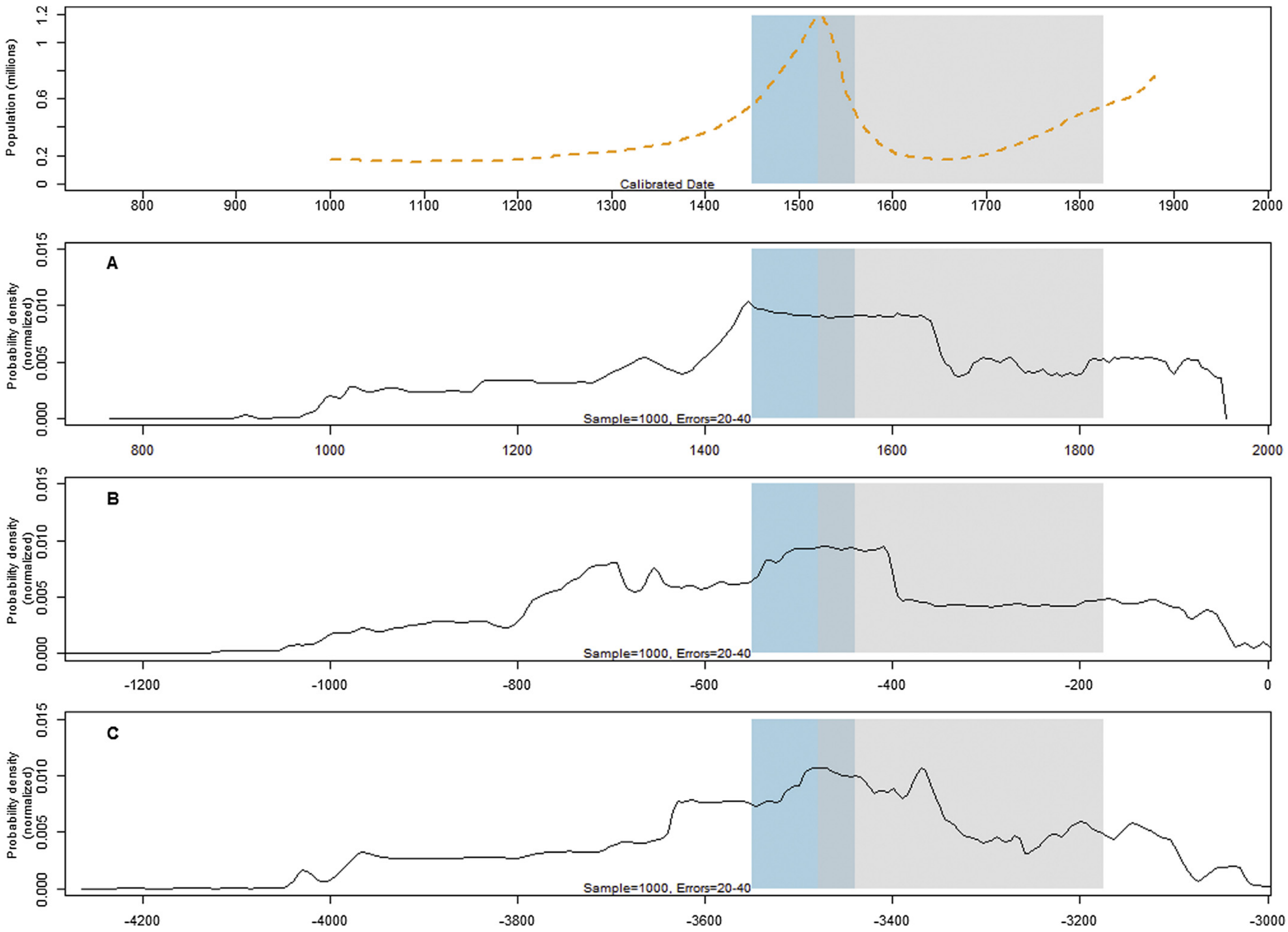

**Fig. 12.** The same set of 1000 simulated dates, set in three different portions of the calibration curve. Both the shape of the curve and the relationship of the curve to the periods of interest (i.e. the spike and trough in the Basin of Mexico population data) vary depending on the region of the calibration curve that the dates occupy.

highlights the difficulty of interpretation of Sum distributions. *Identifying* their occurrence, in the absence of a demographic curve to which Sum proxies may be compared, remains a substantial challenge.

The most fundamental problem of all is that of recognition – i.e., what constitutes an event and what not? In addition to the quality and character of our data, we must be concerned with the magnitude of fluctuation, duration of fluctuation, and even timing of fluctuation. Distinguishing between real and false positives and negatives, in the absence of a reference curve as we have employed in Figs. 13 and 14, becomes a major challenge for Sum approaches. Deviation from either an ideal distribution or a modelled underlying trend (e.g., Shennan et al., 2013) is one possibility whose use has been explored, but such approaches still face the problem of distinguishing the duration and the magnitude of such deviations, as well as determining whether population change or some confounding factor has produced the fluctuation(s).

### 4.5.1. *False positives and negatives*

One risk is that peaks or troughs that result from random sampling or the calibration curve might be interpreted as reflecting population changes. In these examples (Figs. 13A and 14B), a prominent peak or trough appears in the Sum distribution in a period in which the original population is uniform (or trending in the opposite direction).

The inverse – erasure of real patterns by random sampling, calibration, or measurement scatter – is also a risk. In such cases, the Sum distribution does not show a peak or trough where one should be, based on the underlying population, or at least, the deviation from uniformity is no more prominent than other deviations that are clearly spurious (Fig. 13B; 13A is arguable at best).

### 4.5.2. *Timing*

Even in cases when a peak or trough may be accurately and confidently detected, it can be difficult to date it correctly, or to estimate its duration (Figs. 13 and 14; magnitudes of timing errors approximately 70–130 years). This can potentially create false synchronisms with precisely-dated events, or disguise real synchronisms.

### 4.5.3. *Scale*

The amplitude of a peak or trough depends on many factors, and it is difficult (if indeed possible) to translate this into a proportional change in population, or to set appropriate thresholds for significance. In addition, we have to define models of what constitutes a demographic crisis and re-population, or a population boom, to be able to specify the spikes we are looking for. We know, for example, that the Basin of Mexico population fell by 90% between 1520 and 1600, but we only see a drop of approximately 50%, about 100 years later, in both the Sum distributions in Fig. 14. Any other decline

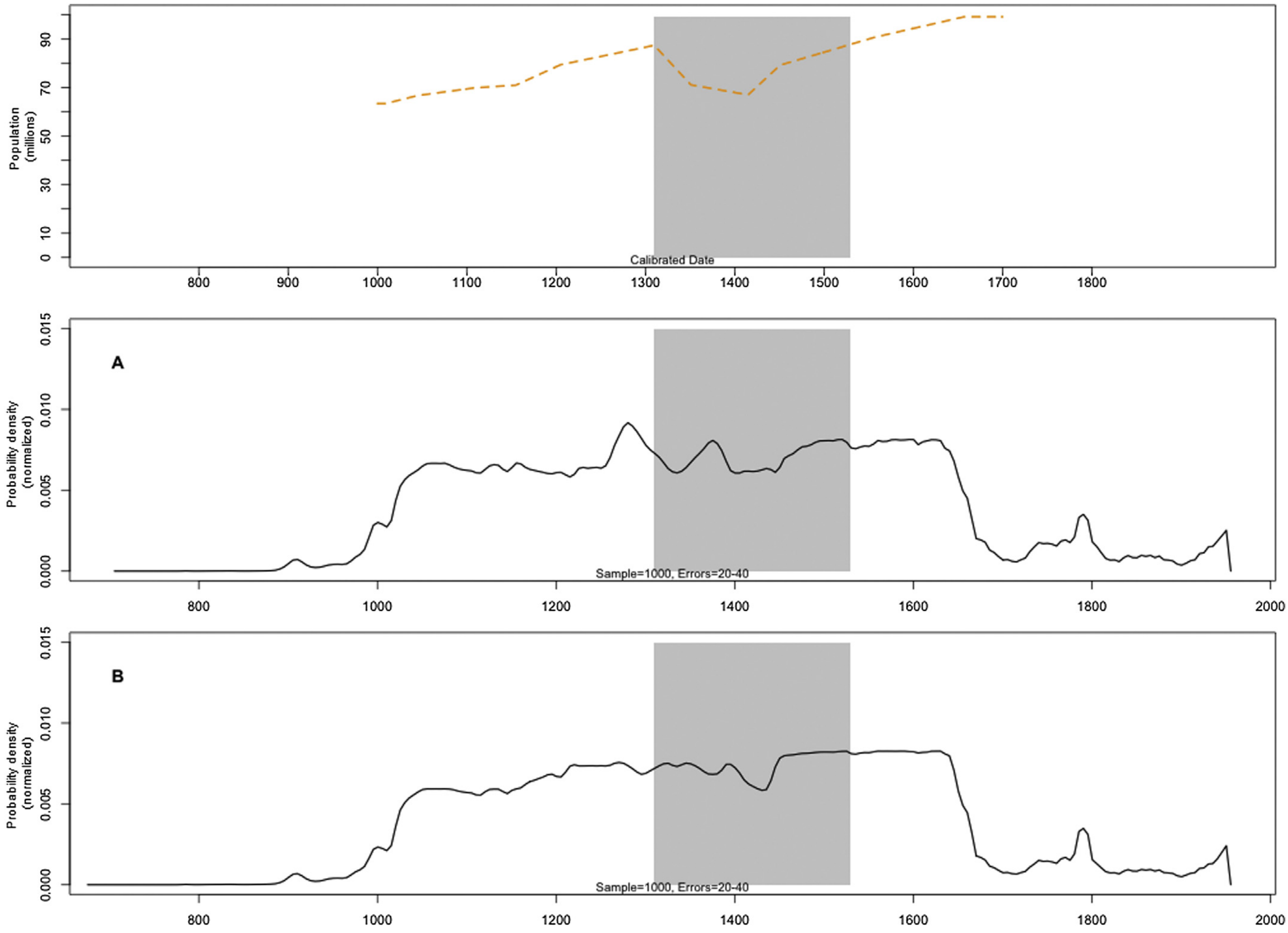

**Fig. 13.** Two runs of 1000 samples from the Black Death produced with *R_Simulate*. The upper run (A) shows a probable true positive in the form of a minimum at about cal AD 1350, a false positive in the spike just before cal AD 1400, and puts the timing of the population crash somewhat too early. The lower run (B) shows no crash – a false negative – at cal AD 1350, and a trough at cal AD ~1450 that may be either a false positive or a mis-timed true positive.

visible in Fig. 14 is spurious, by definition, yet (by comparison with Fig. 13) the 14th-century "false positive" declines could also correspond to events more devastating than the Black Death.

*4.5.4. Edge effects*

One of the effects of measurement uncertainty and calibration is to spread the calibrated $^{14}$C dates beyond the range of time from which samples were drawn – e.g., in Fig. 3 samples from cal AD 1000–1700 produce a Sum distribution that spans approximately the period cal AD 800–1950. These edge effects are obvious in $^{14}$C samples simulated from a known population, but present a problem when dealing with an unknown archaeological population that is itself the object of investigation, particularly if researchers are primarily interested in the start or end of settlement (e.g., Rieth et al., 2011; Mulrooney, 2013) and are therefore unable to ignore the tails of their Sum distributions. In other situations, authors have deliberately summed dates from a wider timespan than the period of interest, in order to minimize edge effects, but we are not aware of any rigorous testing of this method.

## 5. Concluding remarks

The simulations that we have discussed, of historically attested population fluctuations with significant social/political repercussions, show that even under ideal conditions, it is difficult to distinguish between real and spurious population patterns, or to accurately date sharp fluctuations, even with data densities much higher than in most published attempts. Both advocates and critics of a 'Sum' approach might hope that simulation studies would produce baseline criteria regarding both requisite data densities and the duration and magnitude of posited population fluctuations – i.e., a simple means of estimating whether a fluctuation is of a magnitude and duration sufficient to suggest that it reflects real population patterning rather than statistical noise. Unfortunately, the abundance of variables – e.g., population size, event duration and magnitude, magnitudes of dating errors, calibrated date range – and the irregular population distributions that we are trying to reconstruct (of interest precisely because they are irregular – i.e., population crashes or spikes), mean that any such rule of thumb is likely to represent wishful thinking rather than rigorous evaluation. In addition, judgements about the 'reality' of patterns detected in summed radiocarbon data face the same challenge identified by Cowgill (1977) with regard to significance tests in archaeology: their proper use requires judgement and argument rather than binary acceptance/rejection.

While it is possible that large numbers of truly random samples, whose frequency of production, preservation, and analysis is proportional to population size (and not affected by factors such as differential preservation, variable research intensity, etc.), may produce summed calibrated $^{14}$C probability distributions that

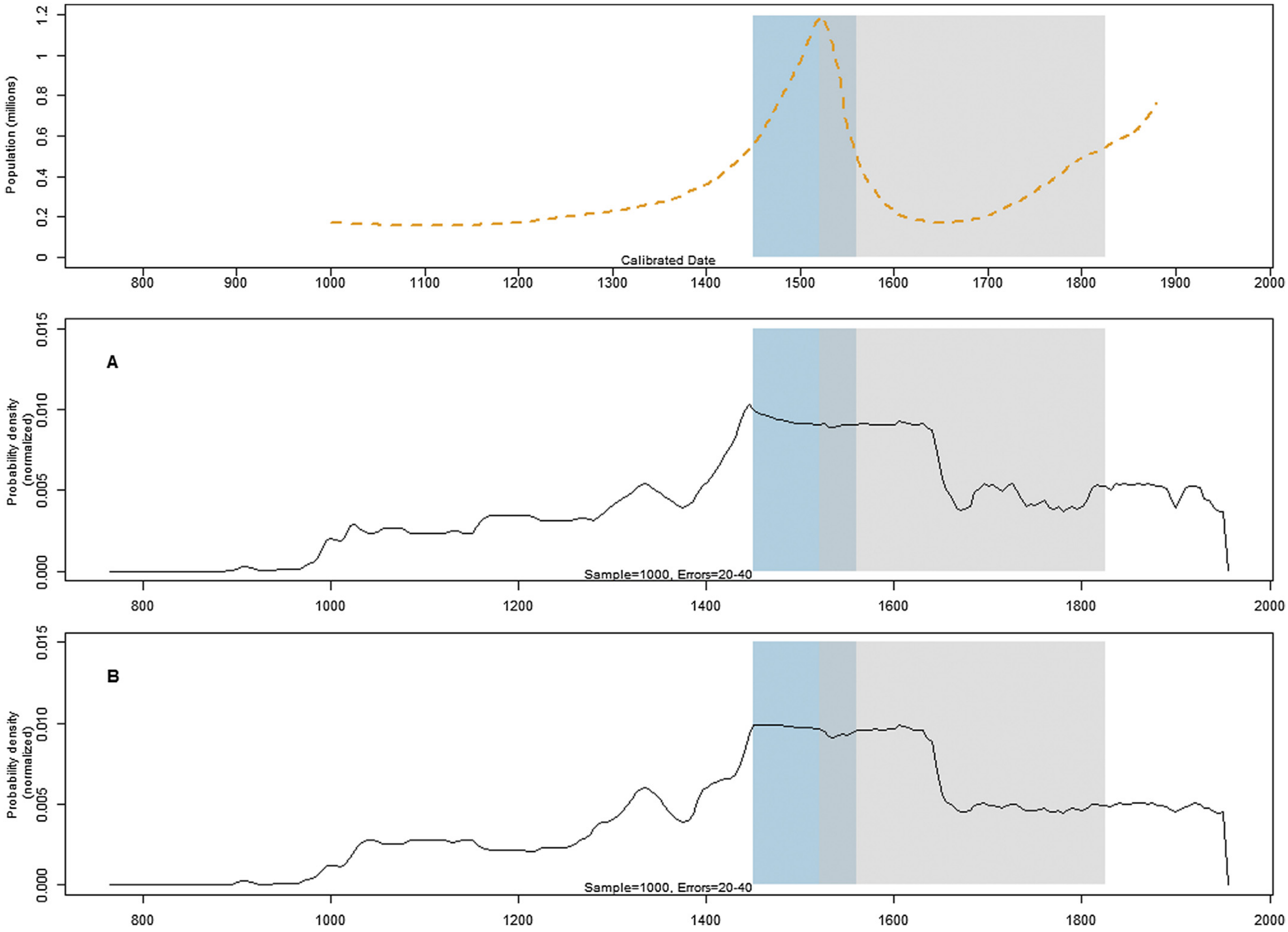

**Fig. 14.** Two runs of 1000 samples from the Basin of Mexico produced with *R_Simulate*. Both runs show the real positive of the Basin of Mexico population spike, but peaking about cal AD 1450 rather than cal AD 1520; similarly both detect a crash, but not occurring until about cal AD 1650. In other words, timing is the chief problem here, with the related problem that the spike appears more a plateau than a peak. Both examples also show false positives in the 14th century, apparently artifacts of the calibration curve, as they appear in both data sets.

identify population fluctuations in prehistory, our simulations clearly indicate that extreme caution is warranted. We have demonstrated that the Sum approach can obscure, as well as reveal, rapid population changes, and suspect that the potential for false positives and mis-timed events could suggest misleading links between population and exogenous events.

In this paper, we have focussed on the use of Sum distributions to detect population fluctuations rather than long-term trends, but the `datesim` tool we provide might easily also be used to examine the identification of long-term trends. We have used the Black Death and Basin of Mexico illustrations because archaeologists are and should be interested in population fluctuations on centennial or shorter timescales; indeed the case studies cited in Table 1, although they in some cases span several millennia, often focus on shorter episodes within these periods, in which the population is perceived to change rapidly (e.g., Kelly et al., 2013; Shennan et al., 2013). As Shennan says, “Given the speed at which demographic processes operate … it is likely that they have a characteristic time scale of 100s rather than 1000s of years (…), so the temporal resolution of our methods must be able to deal with this. A variety of methods for detecting the existence of population fluctuations at the required scale of resolution is now available, not least the use of summed radiocarbon date distributions as a demographic proxy.” (Shennan, 2013:302).

This is not to say that longer-term trends may not also be of interest, and some of the same issues will also arise in studies of population trajectories over timescales of many millennia (e.g., Peros et al., 2010; Williams, 2012). In these cases too a simulation approach is likely to be useful in assessing the relative effects of measurement scatter, calibration effects, taphonomy, and research bias; our results here suggest that while the former two may be less significant in obscuring long-term trends in Sum distributions, effects stemming from differential preservation and research intensity may be more prominent. Moreover, studies of long-term trends in Sum distributions are of use only to the degree to which they can improve on the chronological resolution available from other archaeological proxies for population trends (e.g. in many cases, the number of sites occupied per millennium can already be estimated from artifact assemblages; Tallavaara et al. (2010) even propose a Ceramic Site Frequency Index covering 5000–500 BC with 100-year resolution).

In the case of either short-term fluctuations or long-term trends, we would recommend on the basis of our results that in any situation where the Sum approach is applied, it is incumbent upon researchers to argue 1) that the link between production, preservation, and analysis of datable organic material and population is *in that case* sound, 2) that the $^{14}$C dates they employ can be considered a random sample, 3) that the event(s) they seek to identify are of sufficient duration, relative to the average measurement uncertainties, that it will be possible to identify them, 4) that the sample size is large enough, relative to the span of time under consideration, to be able to identify events of the magnitude postulated, and 5) that the observed patterns are *not* the product of the calibration curve. As problems 3, 4, and 5 are directly related to the posited prehistoric population distribution – itself often of interest precisely because it is irregular, i.e., does not conform to a uniform or other standard distribution – they are best investigated via simulation approaches such as the use of the `datesim` tool presented here. Approaches that seek to test null hypotheses which posit that the observed patterns in a Sum distribution could have been produced by randomly sampling a simple population curve (e.g., representing a constant population) focus on identifying which departures from the null-hypothesis curve are statistically significant; the results are assertions about past population fluctuations (e.g. Shennan et al., 2013:Table 1) that could be tested using the simulation tool that we present here. Only if the same pattern is clearly and consistently reproduced by Sum distributions of simulated dates with comparable data density and measurement uncertainties to the archaeological $^{14}$C results, drawn randomly with probabilities based on the proposed population trajectory, may it be argued that a $^{14}$C sum represents population fluctuations rather than factors such as random sampling, the calibration curve, research intensity, and taphonomic processes.

Minimally, therefore, researchers would be well-advised to investigate via simulation whether it is in fact possible, much less likely, to detect a population event of the magnitude and duration that they propose to identify, given the numbers of samples available and the range of associated measurement errors. If the posited demographic pattern is in principle detectable, it remains of course to argue that in any given case it has been interpreted accurately. We hope the method that we have detailed here can enable such a reflexive approach.

## Acknowledgements

This study was made possible by Daniel Contreras' stay as a Humboldt Fellow at the Institute for Ecosystem Research at CAU Kiel, where he was hosted by Ingmar Unkel. Ian Robertson has been remarkably generous in his help with coding in R. The manuscript has been improved by the helpful comments of two anonymous reviewers.

## References

Antoine, Daniel, 2008. The archaeology of "Plague." Med. Hist. Suppl. 27, 101–114.

Antonson, Hans, 2009. The extent of farm desertion in central Sweden during the late medieval agrarian crisis: landscape as a source. J. Hist. Geogr. 35 (4), 619–641.

Armit, Ian, Swindles, Graeme T., Becker, Katharina, 2013. From dates to demography in later prehistoric Ireland? Experimental approaches to the meta-analysis of large $^{14}$C data-sets. J. Archaeol. Sci. 40 (1), 433–438.

Ballenger, Jesse A.M., Mabry, Jonathan B., 2011. Temporal frequency distributions of alluvium in the American Southwest: taphonomic, paleohydraulic, and demographic implications. J. Archaeol. Sci. 38 (6), 1314–1325.

Bamforth, Douglas B., Grund, Brigid, 2012. Radiocarbon calibration curves, summed probability distributions, and early Paleoindian population trends in North America. J. Archaeol. Sci. 39 (6), 1768–1774.

Bayliss, Alex, Bronk Ramsey, Christopher, van der Plicht, Johannes, Whittle, Alasdair, 2007. Bradshaw and Bayes: towards a timetable for the Neolithic. Camb. Archaeol. J. 17 (Suppl. S1), 1–28.

Bayliss, Alex, Hines, John, Høilund Nielsen, Karen, McCormac, Gerry, Scull, Christopher, 2013. Anglo-Saxon Graves and Grave Goods of the 6th and 7th Centuries AD: a Chronological Framework. Society for Medieval Archaeology Monographs, pp. 340–345.

Bennett, M.K., 1954. The World's Food. Harper & Bros., New York.

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

Bronk Ramsey, Christopher, 2001. Development of the radiocarbon calibration program. Radiocarbon 43 (2; Part A), 355–364.

Bronk Ramsey, Christopher, 2009a. Bayesian analysis of radiocarbon dates. Radiocarbon 51 (1), 337–360.

Bronk Ramsey, Christopher, 2009b. Dealing with outliers and offsets in radiocarbon dating. Radiocarbon 51 (3), 1023–1045.

Buchanan, Briggs, Collard, Mark, Edinborough, Kevan, 2008. Paleoindian demography and the extraterrestrial impact hypothesis. Proc. Natl. Acad. Sci. U. S. A. 105 (33), 11651–11654.

Buchanan, Briggs, Hamilton, Marcus, Edinborough, Kevan, O'Brien, Michael J., Collard, Mark, 2011. A comment on Steele's (2010) "Radiocarbon dates as data: quantitative strategies for estimating colonization front speeds and event densities." J. Archaeol. Sci. 38 (9), 2116–2122.

Chiverrell, Richard C., Thorndycraft, Varyl R., Hoffmann, Thomas O., 2011. Cumulative probability functions and their role in evaluating the chronology of geomorphological events during the Holocene. J. Quat. Sci. 26 (1), 76–85.

Collard, Mark, Edinborough, Kevan, Shennan, Stephen, Thomas, Mark G., 2010. Radiocarbon evidence indicates that migrants introduced farming to Britain. J. Archaeol. Sci. 37 (4), 866–870.

Cowgill, George L., 1977. The trouble with significance tests and what we can do about it. Am. Antiq. 42 (3), 350–368.

Crombé, Philippe, Robinson, Erick, 2014. $^{14}$C Dates as demographic proxies in Neolithisation models of northwestern Europe: a critical assessment using Belgium and northeast France as a case-study. J. Archaeol. Sci. 52, 558–566. http://dx.doi.org/10.1016/j.jas.2014.02.001.

Culleton, Brendan J., 2008. Crude demographic proxy reveals nothing about Paleoindian population. Proc. Natl. Acad. Sci. U. S. A. 105 (50), E111.

Drennan, Robert D., 2009. Statistics for Archaeologists: a Common Sense Approach. Springer.

Durand, John D., 1977. Historical estimates of world population: an evaluation. Popul. Dev. Rev. 3 (3), 253–296.

Gamble, Clive, Davies, William, Pettitt, Paul, Hazelwood, Lee, Richards, Martin, 2005. The archaeological and genetic foundations of the European population during the Late Glacial: implications for "Agricultural thinking." Camb. Archaeol. J. 15 (02), 193–223.

Gkiasta, Marina, Russell, Thembi, Shennan, Stephen, Steele, James, 2003. Neolithic transition in Europe: the radiocarbon record revisited. Antiquity 77 (295), 45–62.

Hinz, Martin, Feeser, Ingo, Sjögren, Karl-Göran, Müller, Johannes, 2012. Demography and the intensity of cultural activities: an evaluation of Funnel Beaker Societies (4200–2800 cal BC). J. Archaeol. Sci. 39 (10), 3331–3340.

Hoffmann, Thomas, Lang, Andreas, Dikau, Richard, 2008. Holocene river activity: analysing $^{14}$C-dated fluvial and colluvial sediments from Germany. Quat. Sci. Rev. 27 (21–22), 2031–2040.

Johnstone, Eric, Macklin, Mark G., Lewin, John, 2006. The development and application of a database of radiocarbon-dated Holocene fluvial deposits in Great Britain. Catena 66 (1–2), 14–23.

Kelly, Robert L., Surovell, Todd A., Shuman, Bryan N., Smith, Geoffrey M., 2013. A continuous climatic impact on Holocene human population in the Rocky Mountains. Proc. Natl. Acad. Sci. U. S. A. 110 (2), 443–447.

Kerr, T.R., McCormick, F., 2014. Statistics, sunspots and settlement: influences on sum of probability curves. J. Archaeol. Sci. 41, 493–501.

Livi-Bacci, Massimo, 1999. The Population of Europe. Blackwell, Oxford.

Livi-Bacci, Massimo, 2006. The depopulation of Hispanic America after the Conquest. Popul. Dev. Rev. 32 (2), 199–232.

Lovell, W George, 1992. "Heavy shadows and black night": disease and depopulation in colonial Spanish America. Ann. Assoc. Am. Geogr. 82 (3), 426–443.

McCaa, Robert, 1995. Spanish and Nahuatl views on smallpox and demographic catastrophe in Mexico. J. Interdiscip. Hist. 25 (3), 397–431.

Mulrooney, Mara A., 2013. An island-wide assessment of the chronology of settlement and land use on Rapa Nui (Easter Island) based on radiocarbon data. J. Archaeol. Sci. 40 (12), 4377–4399.

Pamuk, S., 2007. The Black Death and the origins of the "Great Divergence" across Europe, 1300–1600. Eur. Rev. Econ. Hist. 11 (3), 289–317.

Parsons, Jeffrey R., 1974. The development of a prehistoric complex society: a regional perspective from the Valley of Mexico. J. Field Archaeol. 1 (1), 81–108.

Peros, Matthew C., Munoz, Samuel E., Gajewski, Konrad, Viau, André E., 2010. Prehistoric demography of North America inferred from radiocarbon data. J. Archaeol. Sci. 37 (3), 656–664.

Prates, L., Politis, G., Steele, J., 2013. Radiocarbon chronology of the early human occupation of Argentina. Quat. Int. 301, 104–122.

R. Core Team, 2013. R: a Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Reimer, Paula J., Bard, Edouard, Bayliss, Alex, Beck, J Warren, Blackwell, Paul G., Bronk Ramsey, Christopher, Buck, Caitlin E., Cheng, Hai, Edwards, R Lawrence, Friedrich, Michael, Grootes Pieter, M., Guilderson Thomas, P., Haflidason Haflidi, Hajdas Irka, Hatté Christine, Heaton Timothy, J., Hoffmann Dirk, L., Hogg Alan, G., Hughen Konrad, A., Kaiser, K., Felix, Kromer Bernd, Manning Sturt, W., Niu Mu, Reimer Ron, W., Richards David, A., Scott, E., Marian, Southon John, R., Staff Richard, A., Turney Christian, S.M., van der Plicht Johannes, 2013. IntCal13 and Marine13 radiocarbon age calibration curves 0–50,000 years cal BP. Radiocarbon 55 (4), 1869–1887.

Rick, John W., 1987. Dates as data: an examination of the Peruvian preceramic radiocarbon record. Am. Antiq. 52 (1), 55–73.

Riede, Felix, 2009. Climate and demography in Early Prehistory: using calibrated $^{14}$C dates as population proxies. Hum. Biol. 81 (2–3), 309–337.

Riede Felix, 2008. The Laacher See-eruption (12,920 BP) and material culture change at the end of the Allerød in Northern Europe. J. Archaeol. Sci. 35 (3), 591–599.

Rieth, Timothy M., Hunt, Terry L., Lipo, Carl, Wilmshurst, Janet M., 2011. The 13th century polynesian colonization of Hawai'i Island. J. Archaeol. Sci. 38 (10), 2740–2749.

Ruddiman, William F., 2003. The anthropogenic greenhouse era began thousands of years ago. Clim. Change 61 (3), 261–293.

Shennan, Stephen, 2009. Evolutionary demography and the population history of the European Early Neolithic. Hum. Biol. 81 (2–3), 339–355.

Shennan, Stephen, 2013. Demographic continuities and discontinuities in Neolithic Europe: evidence, methods and implications. J. Archaeol. Method Theory 20 (2), 300–311.

Shennan, Stephen, Edinborough, Kevan, 2007. Prehistoric population history: from the Late Glacial to the Late Neolithic in Central and Northern Europe. J. Archaeol. Sci. 34 (8), 1339–1345.

Shennan, Stephen, Timpson, Adrian, Edinborough, Kevan, Colledge, Susan M., Kerig, Tim, Manning, Katie, Thomas, Mark G., Downey, Sean S., 2013. Regional population collapse followed initial agriculture booms in mid-Holocene Europe. Nat. Commun. 4, 1–8.

Spriggs, Matthew, 1989. The dating of the Island Southeast Asian Neolithic: an attempt at chronometric hygiene and linguistic correlation. Antiquity 63 (240), 587–613.

Stuiver, M., Reimer, P.J., 1986–2014. Calib Radiocarbon Calibration Program.

Surovell, Todd A., Brantingham, P Jeffrey, 2007. A note on the use of temporal frequency distributions in studies of prehistoric demography. J. Archaeol. Sci. 34 (11), 1868–1877.

Tallavaara, Miikka, Pesonen, Petro, Oinonen, Markku, 2010. Prehistoric population history in eastern Fennoscandia. J. Archaeol. Sci. 37 (2), 251–260.

Weninger, Bernhard, Jöris, Olaf, Danzeglocke, Uwe, 2007. Cologne Radiocarbon Calibration & Palaeoclimate Research Package.

Whitehouse, Nicki J., Schulting, Rick J., McClatchie, Meriel, Barratt, Phil, Rowan McLaughlin, T., Bogaard, Amy, Colledge, Susan M., Marchant, Rob, Gaffrey, Joanne, Jane Bunting, M., 2013. Neolithic agriculture on the European western frontier: the boom and bust of early farming in Ireland. J. Archaeol. Sci., 1–63.

Williams, Alan N., 2012. The use of summed radiocarbon probability distributions in archaeology: a review of methods. J. Archaeol. Sci. 39 (3), 578–589.

Yeloff, Dan, van Geel, Bas, 2007. Abandonment of farmland and vegetation succession following the Eurasian plague pandemic of AD 1347–52. J. Biogeogr. 34 (4), 575–582.