journal of computer applications in archaeology

Hinz, M. 2020. Sensitivity of Radiocarbon Sum Calibration. *Journal of Computer Applications in Archaeology,* 3(1): 238–252. DOI: https://doi.org/10.5334/jcaa.53

CASE STUDY

Sensitivity of Radiocarbon Sum Calibration

Martin Hinz

Sum calibration has become a standard tool for demographic studies, even though the methodology itself is far from uncontroversial. In addition to fundamental methodological criticism, questions are frequently raised about the sample size and data density required to detect large-scale changes in past populations. This article uses a simulation approach to determine the detection probabilities for events of varying intensity and with varying data density. At the same time, the effectiveness of Monte Carlo-based confidence envelopes as a countermeasure against false-positive results is tested. The results show that the detection of such events is not unlikely and that the Monte Carlo method is well suited to separate signal and noise. However, the nature of the events already observed in this way demands further assessment.

Keywords: Prehistoric demography; Summed radiocarbon date distributions; Simulation; Calibration; Population proxies; Reproducible Research

Introduction

In recent years, the use of more or less large collections of ^{14}C data has almost become a standard tool to estimate demographic developments of the past. The original approach of Rick (1987) assumes that 'if archaeologists recovered and dated a random, known percentage of the carbon from a perfectly preserved carbon deposit to which each person-year of occupation contributed an equal and known amount, they could estimate the number of people who inhabited a region during a given period' (Rick 1987: 56).

There is a history of debate on the use of radiocarbon Summed Probability Distributions (SPDs). Several authors use the method in different elaborations to identify past processes, most often demographic processes (e.g., Armit, Swindles and Becker 2013; Buchanan, Collard and Edinborough 2008; Collard et al. 2013; Gamble et al. 2005; Gkiasta et al. 2003; Hinz et al. 2012; Hoffmann, Lang and Dikau 2008; Johnstone, Macklin and Lewin 2006; Kelly et al. 2013; Mulrooney 2013; Rick 1987; Riede 2009; Rieth et al. 2011; Shennan 2009, 2012; Shennan and Edinborough 2007; Tallavaara, Pesonen and Oinonen 2010; Timpson et al. 2014; Whitehouse et al. 2014). Others reject the method in general, criticise certain aspects, or point out weaknesses (Ballenger and Mabry 2011; Bamforth and Grund 2012; Bayliss et al. 2007; Bleicher 2013; Chiverrell, Thorndycraft and Hoffmann 2011; Contreras and Meadows 2014; Crombé and Robinson 2014; Culleton 2008; Prates, Politis and Steele 2013; Steele 2010; Surovell and Brantingham 2007; Surovell et al. 2009; Torfing 2015; Williams 2012). A critical assessment is very helpful for improving a method or identifying its shortcomings. However, what many critics do not emphasize, and many supporters do not sufficiently consider, is that already in Ricks original paper (Rick 1987: 57–59; **Figure 1**) the essential sources of error (Intervening, Creation, Preservation, and Investigation biases) have been identified.

Due to the widespread use of the methodology in recent years, it is essential to explore the conditions for the meaningful use of sum calibration. Ideally, a catalogue of prerequisites and methodological requirements should be compiled providing a comparable and high standard for the usage of this estimator and a quantification of the uncertainty in its application.

This paper examines some of the objections with the help of simulations. This relates in particular to the methodological questions related to the sensitivity of the method as put forward by Contreras and Meadows (2014). Similar to that paper, it is examined whether ^{14}C summations can be used to identify patterns that can be related to population fluctuations in the past. Thereby it is assumed that the amount of archaeological and thus datable material reflects those changes in particular. The sampling bias and its effects will be addressed. For this purpose, simulated and thus artificial ^{14}C dates are generated, drawing from a probability curve based on a historical event – the Black Death. For different data densities (average number of samples per year), a random sample of years is drawn from the period in question. The probability of each year corresponds to the relative demographic trend for the same year. This means, that the probability of drawing a sample from a particular date is exactly proportional to the population size in that year. The data simulated in this way is treated using Oxcal in the same way as the data is processed in existing studies (i.e. using Oxcal's `Sum` command). It is then checked whether the given patterns can

Institut für Archäologische Wissenschaften, Universität Bern, CH
martin.hinz@iaw.unibe.ch

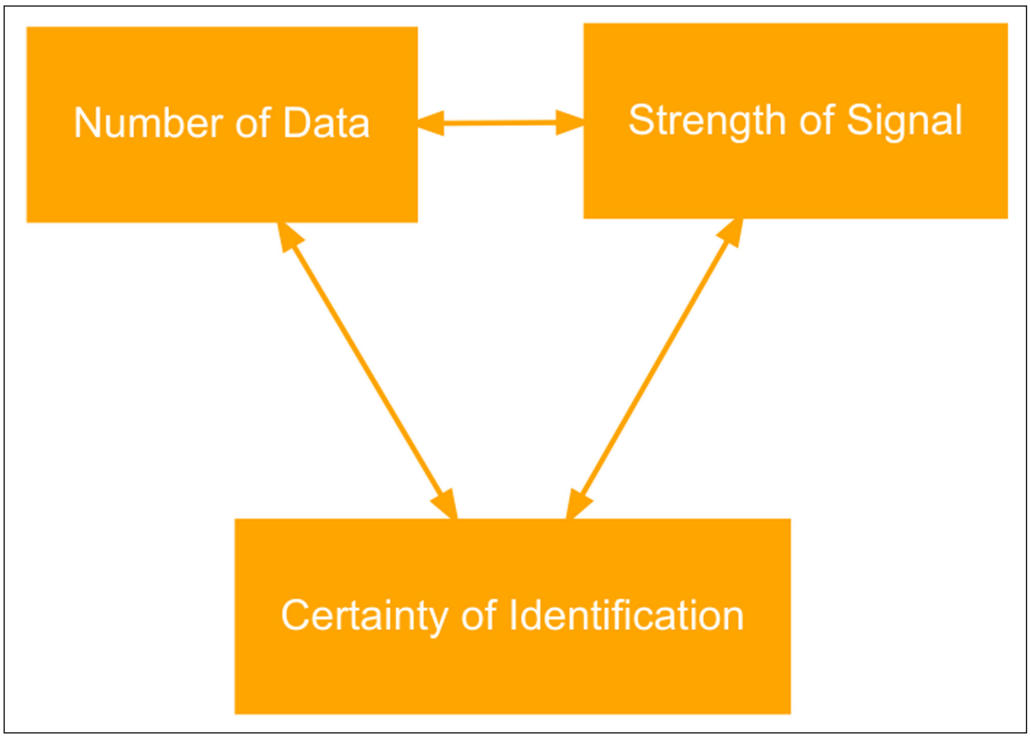

Figure 1: Interdependency between amount of information, intensity of pattern and desired uncertainty.

be found in the calibration results. It should be noted that, as in the above-mentioned study, no further measures against other sources of error apart from the sampling bias are taken. Above all, no binning is used to standardise the data per site, which has now been established as a standard procedure. On the one hand, such an error does not exist in the simulated data, on the other hand, the method used should be as close as possible to that of the paper by Contreras and Meadows (2014). The main focus is on enriching the unquantified results of that study with a quantification of the detection probability.

Background

The potential of using SPDs to assert a statement about past (perhaps demographic) processes depends on three factors (see **Figure 1**):

- Amount of information available (number of data points)
- Intensity of the process to be identified (strength of the signal in the demographic fluctuation)
- Certainty with which this signal is to be identified (or other, permitted uncertainty)

If a very strong signal is to be detected, less data may be sufficient to be able to identify it with a specified uncertainty. Conversely, increased certainty about the validity of the signal requires either more data or a stronger signal. This means that strong demographic fluctuations in the past can be detected with greater certainty even based on smaller amounts of ^{14}C datings, whereby more data would be required in the case of weaker fluctuations. These relationships are fundamental to any kind of statistical hypothesis test.

In their article, Contreras and Meadows (2014) worked on this question. They put all other possible methodological problems to one side (although they did elaborate on them in detail) and investigated how well simulated demographic changes can be tracked by their effect on simulated ^{14}C data. The paper went much further than others due to its simulation approach and its results are therefore largely transparent. This is a very valuable and useful contribution to the debate. Their main case study is the Black Death, whose demographic influence can be adequately understood from written sources. They describe their demographic example as: '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. AD1550' (Contreras and Meadows 2014: 596; cf. also **Figure 2**) However, the details of the setting are irrelevant for the specific investigation, it could have been any arbitrary or randomly generated example. The chosen one serves the authors above all to show that such a devastating event as the Black Death could remain undiscovered by ^{14}C summations.

When using an over-representatively high number of data, or such a data density, with 1000 data points for a period of 1000–1700 BCE (density 1.43 data per year), the Black Death would basically emerge (Contreras and Meadows 2014: Figure 3). However, they argue that the strength of the event in the resulting signal could not be attributed to such a disaster without prior knowledge (Contreras and Meadows 2014: 599). At a density they consider to be closer to the archaeological reality – the authors assume this to be 0.29, representing 200 data for 700 years – the sampling effect would prevent the underlying demographic processes from being properly represented by the simulated ^{14}C data (Contreras and Meadows

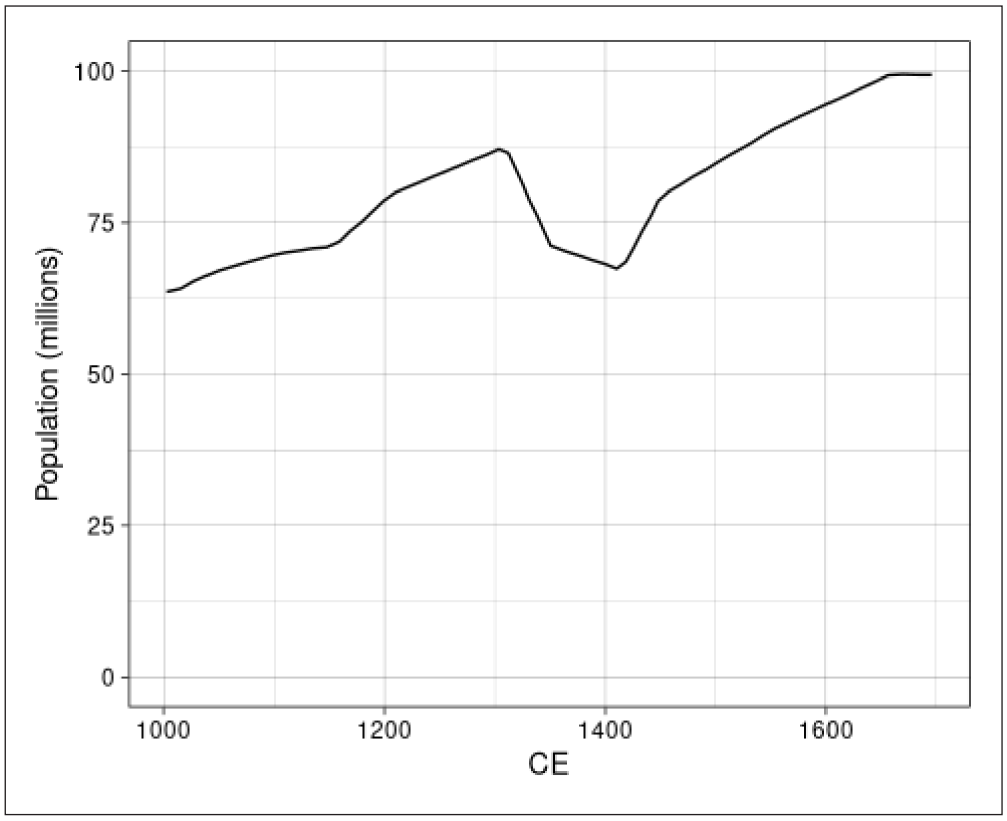

Figure 2: Population development during the time of the Black Plague, according to Contreras and Meadows (2014).

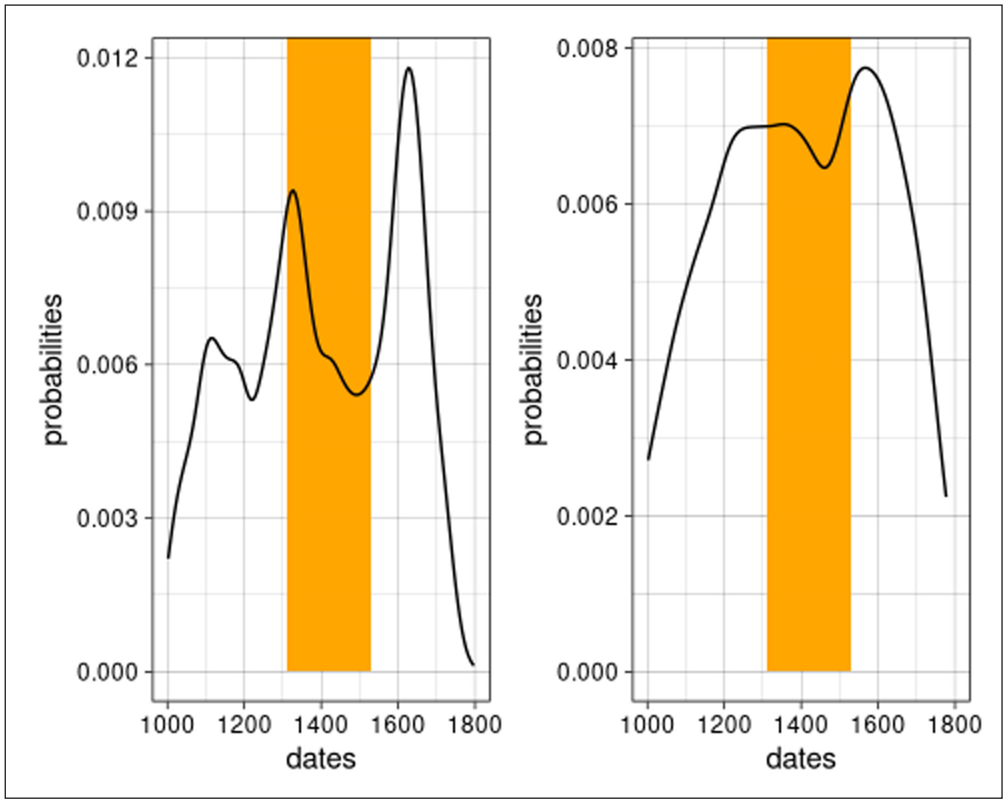

Figure 3: Comparing the same result of a random sum calibration unsmoothed (left) and smoothed (right) with a two-sided smoothing window of total 500 years.

2014: Figure 6). They write: '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' (Contreras and Meadows 2014: 601). In general, the data

density is decisive for the effectiveness of this estimator, whereby even with the maximum simulated number of dates (2000) the Black Death is 'far from obvious' as an event (Contreras and Meadows 2014: 602). In addition, they argue, the temporal fixation of the event is problematic due to the scatter effect especially of legacy data with high standard deviation. Thus it would be not possible to separate signal from noise, to separate false-positive and false-negative from real results, and to identify the exact timing and magnitude of the underlying phenomenon (Contreras and Meadows 2014: 603–605). In their concluding remarks they consequently state 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' (Contreras and Meadows 2014: 605).

With all the importance that the simulation approach adds to this paper, unfortunately, the authors do not use its full potential. Although they created different scenarios of data density, each is only examined with five simulation runs (for 200, 1000 and 2000 samples respectively; Contreras and Meadows 2014: 596). Even if five is more than one, this certainly does not represent a statistically reliable basis for a far-reaching statement. In addition they state, as paraphrased above, that the Black Death could have remained undetected, without further specification or quantification. A significantly higher number of simulations might be mandatory for such a statement. A very important step in this direction has already been taken by McLaughlin (2019), who reviewed the Black Death scenario in his article on using the KDE model for similar analyses, and who has already come up with detection rates. A perfect pattern recognition was achieved with a sample number of 3000. Here, however, only 30 simulation runs were checked in each case, and the effect strength was not varied.

Precisely against this background the triangle of effect strength, data quantity and certainty of identification should be quantified here. Using the same basic pattern, the Black Death, the aim is to determine, for different scenarios of effect strength and data quantity, in how many cases such a demographic catastrophe could have remained undetected. It is primarily a question of false-negative results. False positives can be meaningfully detected by other simulation approaches, as has been discussed elsewhere (e.g., Shennan et al. 2013; Edinborough et al. 2017). It will be applied in a later step (see below).

Methods

The overall approach and the implemented workflow consists of three main parts:

1. The simulation of the ^{14}C data from the underlying population curve,
2. the identification of the signal from the resulting summation curve, and
3. the combination of the results from the individual simulation runs.

To simulate different densities of ^{14}C dates, 18 scenarios were created (30–90 in steps of ten, 100–900 in steps of one hundred, 1000–2000 in steps of one thousand). For each scenario, 200 simulation runs were used. The whole process is controlled by a superimposed control structure.

In the first part of the analysis, the original scenario of Contreras and Meadows (2014) was reconstructed. The population curve was reconstructed and for different numbers of simulated samples, the signal was detected as described below. This process was repeated 200 times for each parameterization of the number of samples in order to obtain a statistical basis for the evaluation. The proportion of detected patterns was recorded, and the scenarios themselves were repeated 200 times to capture the range of variation between runs. Although the scattering of the detection results with respect to the standard deviation of the successful detection is primarily a function of the sample size (200 repetitions) and the true detection rate, this exhaustive test setup was chosen in order to account for any nonlinear effects resulting from the shape of the calibration curve. This resulted in 720,000 individual simulation runs (200 batches of 200 simulations of 18 scenarios).

The signal strength, i.e. the intensity with which the demographic signal decreases, is 77.4% in the 'real' data of the Black Death. In the second part of the analysis, signal strengths of 30%–90% were simulated in steps of ten, respectively the data set of the Black Death was changed in such a way that such a demographic change is predetermined by the data set. This results in a total of 126 scenarios. For each of the scenarios, 200 simulation runs were carried out, resulting in a total of 25,200 individual runs. The repetition of individual scenarios was omitted as this would have considerably increased the runtime of the algorithm.

This process was repeated for both settings including the test against false positives as described below. In total, the whole simulation includes 1,490,400 individual sum calibrations. The choice for the final number of runs and repetitions resulted from the total run time, which was 94,480 seconds or 26 hours and 15 minutes (using parallel computing on 6 cores of an Intel(R) Xeon(R) CPU E3-1240 v5 at 3.50GHz with 16 GB RAM).

Simulation of the ^{14}C dates

For the simulation of the ^{14}C data, the original curve from Contreras and Meadows (2014) was used. The data was converted into numerical values by digitizing (using the software Engauge). The corresponding data set is attached as supplementary material or can be accessed in the reproducible analysis.

The population numbers were then interpolated by linear approximation on an annual basis and converted into a probability distribution by normalization to the sum of 1. This distribution then served as weighting for a random drawing of calendar dates representing the individual sample. The sample size was defined as a scenario based on the given parameterisation (see above). In the second part of the analysis, this distribution was changed by parameterising the signal strength by linear rescaling in such a way that the drop from the peak before the

demographic signal to the minimum of the curve corresponds to the given signal strength.

The random years obtained in this way, whose frequency corresponds to the given population curve of the Black Death, were then processed as a sum calibration using `C_Simulate` and `Sum` and calibrated via OxCal (using the package oxcAAR; Hinz et al. 2018). As in Contreras and Meadows' study (2014), the standard deviation was randomly sampled equally distributed in the range of 20–40 years.

The smoothing of the resulting calibration result with a moving average, as suggested by (Williams 2012) with a window of 500 years minimum, was considered, but rejected again. The reason for this is that the more turbulent curve of the calibration result produces a more realistic scenario (see **Figure 3**).

Detection of the signal

To achieve an automated detection of the signal in the calibration result, an algorithm was written which performs this task. The local minima between 1210 and 1630 were recorded and the strongest minimum was selected. If this was not in the period between 1310 and 1530, i.e. the minimum in the population curve of the Black Death, the result was discarded as a non-match. It was then tested whether this minimum was at least 10% below the mean of the 100 years preceding and following the event with a lag of 50 years (1260 resp. 1580). Only if this was the case the signal was considered detected. A selection of random examples of accepted and rejected calibration results can be found in **Figure 4** resp. **5**, or can be easily generated using the reproducible code itself.

Combination of the results

The results of the individual runs were recorded and stored in tabular form. For the first part of the analysis, fixing the signal strength to the value corresponding to the original examination (Contreras and Meadows 2014), the number of detections per run, normalized to the total number of runs, was recorded. For the second part, since only one run with 200 repetitions was performed per scenario, only one value was recorded for each scenario.

Accordingly, mean value, standard deviation, internal quartile and 95% interval can be calculated for the original scenario. For the second part, on the other hand, only one value per scenario is shown, but the influences of sample size and signal strength can be calculated individually.

Elimination of false positive results

Shennan et al. (2013) used a Monte Carlo simulation method that produces simulated data distributions under an adjusted null model. These are then used to test characteristics in the observed data set for statistically significant patterns. A large number of individual simulations are

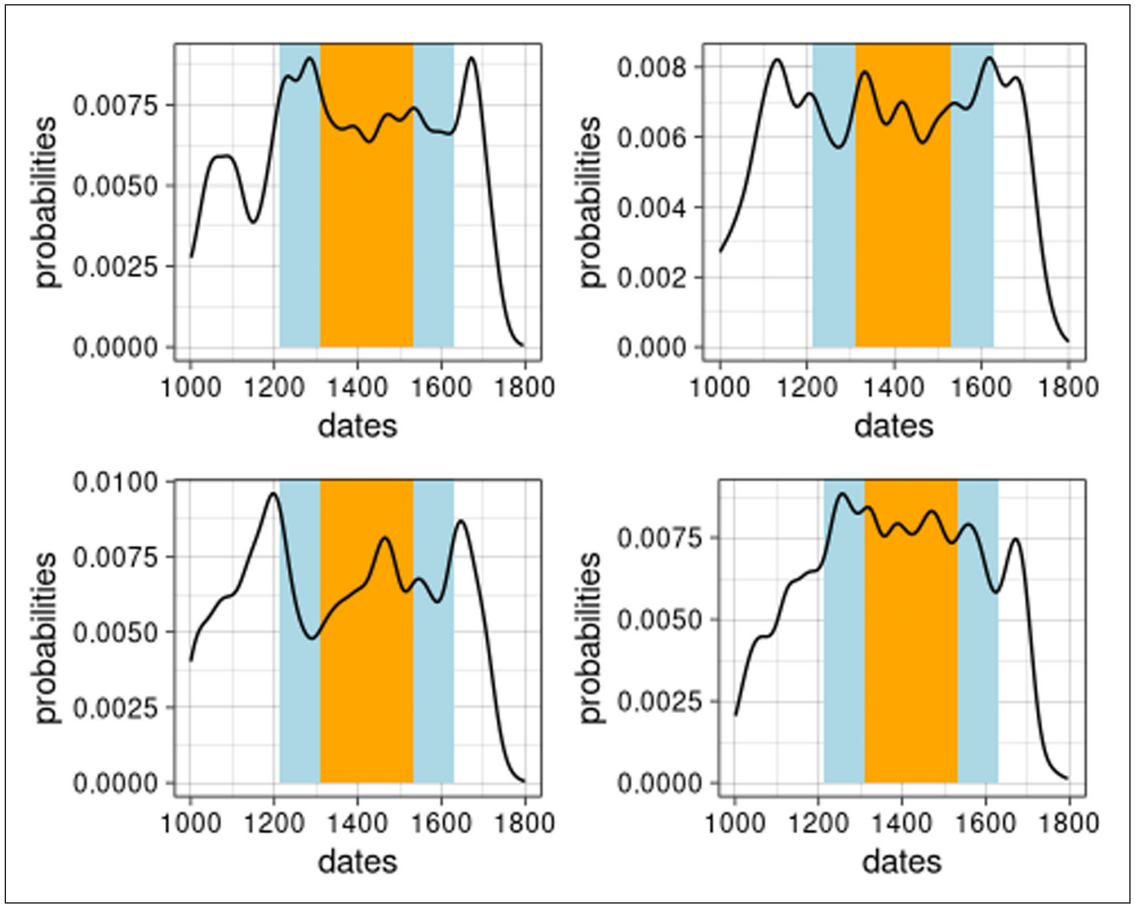

Figure 4: Four examples of rejected results (signal not detected) using the original signal strength and 200 dates. Orange Area: where a minimum should be present. Blue Area: Where the signal should be at least 10% higher than in the minimum on average.

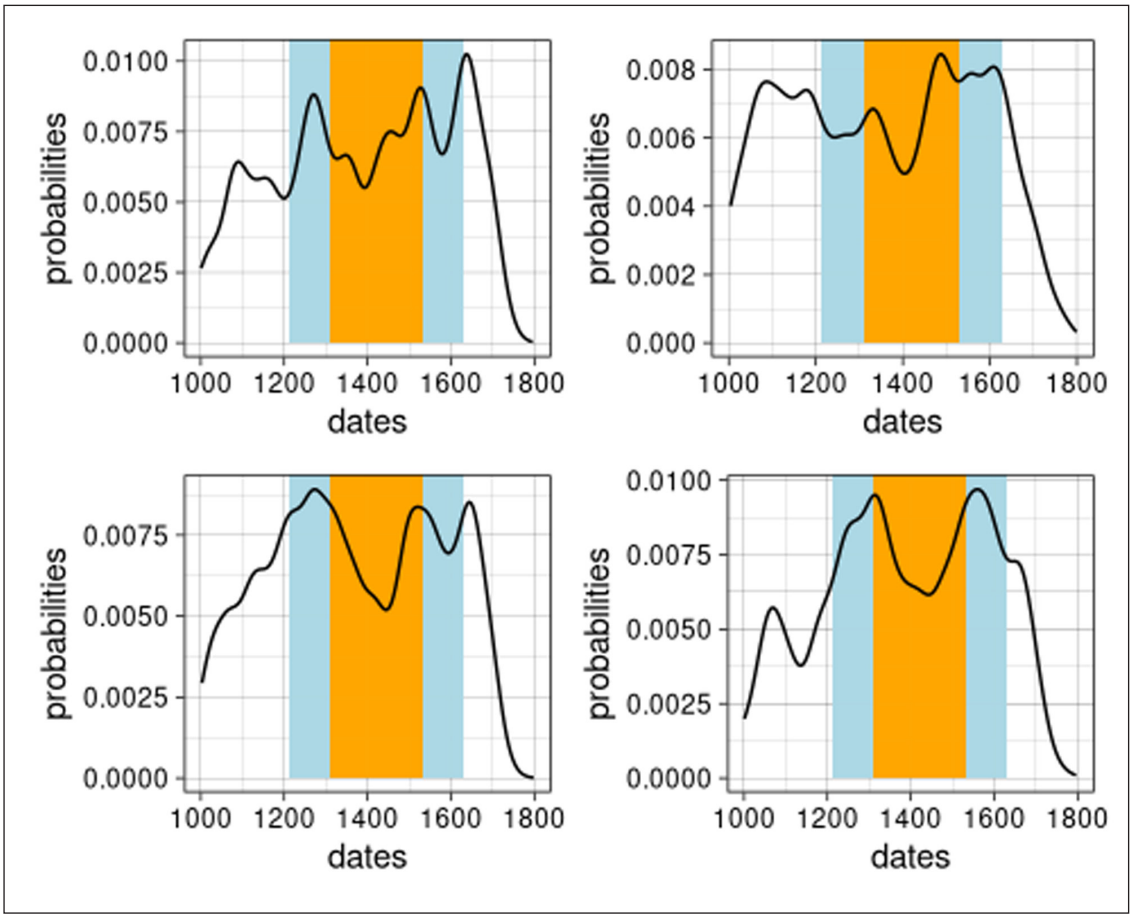

Figure 5: Four examples of accepted results (signal detected) using the original signal strength and 200 dates. Orange Area: where a minimum should be present. Blue Area: Where the signal should be at least 10% higher than in the minimum on average.

carried out using the null model as the population curve, similar to the simulation technique described above. The interval in which the simulated data ranges reflects the element of random sample distribution. Since the 5% significance boundary is set as the statistical standard, the 95% interval (i.e. the quantiles 0.025 and 0.975) is usually taken from the simulated data. A signal, to be evaluated as significant and thus 'real', must lie outside this fluctuation range.

This approach, with slightly different settings, has since become established as the standard procedure for checking the patterns detected in sum calibrations. While, for example, Shennan et al. (2013) uses an exponential generalized linear model for the null model, which is adapted to the data, a simpler approach is chosen here as in other publications (Hinz et al. 2019). The null model is a uniform distribution of the data within a specific time window. Thus, no assumption about a possible population development is made in advance, as would be the case with an exponential function in the sense of population growth. With this, I assume a stable population, and those events, which fall out of the hull generated by the simulation, can be considered as significantly different from this null model. A specific helper function is implemented in the package oxcAAR (Hinz et al. 2018) (`oxcalSumSim()`), which can be used to easily perform such a simulation. It has to be noted that this function is based on `R_Simulate` of OxCal, and therefore shows rather wider uncertainty ranges than it would be necessary for `C_Simulate`. In the given context, this rather increases the robustness of the estimation.

For the original methodology of Shennan et al. (2013) an extension has recently been proposed (Edinborough et al. 2017), that allows a more local and specific approach to hypothesis testing with respect to sum calibration. This expansion will not be further explored in the following, even though it has been successfully applied to the Black Death scenario. The reason is that in this paper I am mainly interested in the general detectability even in the absence of previous knowledge (as it may be available from literary sources), and therefore prefer the simplest possible parameterization.

Reproducible Research in Simulation Studies

Reproducibility has not yet become the standard for archaeological analysis. In many cases, the way archaeological data are collected prevents complete reproducibility of results, as an excavation can only be carried out once. However, in the case of derived, secondary analyses, reproducibility is clearly a preferable design consideration in any research. This is all the more true for simulation studies, which naturally rely on random effects and should therefore be reproducible in their parameterization and which also create the perfect conditions for such a research design regarding their database.

Unfortunately, especially in the field of summed ^{14}C analyses, it is often the case that the argumentation relates on single observations or single calibration runs, i.e. only few results are presented as pars pro toto. At the same time, the source code used to generate these numbers is usually not included in the paper and is also not accessible elsewhere. Therefore the results must be believed as *argumentum ab auctoritate*. A listing of related papers is deliberately omitted here.

If the source code is available or at least reconstructable (as in Contreras and Meadows 2014), a big step towards reproducibility has already been taken. In this article I try to go one step further and choose an Open Science approach in the sense of reproducible research (in the sense of Marwick 2017). The code underlying the simulations is made available together with the article, based on the package rrtools (https://github.com/benmarwick/rrtools). It is available as an R package (sensitivity.sumcal.article.2020) and can be obtained directly (https://github.com/MartinHinz/sensitivity.sumcal.article.2020) or from a repository (Zenodo, doi: 10.5281/zenodo.3613674). With this, all results should be easily reproducible and verifiable, especially the settings of the simulation should be available for direct verification.

Results

Original Setup

The results of the reproduction of the original scenario can be seen in **Table 1**. For the situation of 1000 samples for 700 years described by the authors as super-ideal (results in a density of 1.43) a detection rate of 72.1% results. In half of the cases, the value was between 70% and 74%, 95% of the values lay between 66.5% and 77%.

For the case of a sample size of 200, which Contreras and Meadows (2014: 601) estimated as realistic, the mean detection rate is 68.9%, with the inner quartile between 66.5% and 71% and the 95% interval between 63% and 76%.

Thus, the estimation of Contreras and Meadows (2014) was not completely unjustified. The signal could have been overlooked, following the original simulation setup, with a probability of 1/3. The detection chance seems to be relatively independent of the sample size (once the sample density has surpassed 300).

This can be seen in **Table 1**, more clearly perhaps from the box plot of the results (**Figure 6**) or the representation as regression (with logarithmic x-axis, **Figure 7**). Up to a sample size of about 300, corresponding to a density of 0.43 dates per year, the detection rate improves and then remains on a plateau.

The results indicate that, on the one hand, there is a clear chance to detect an event like the Black Death with a tool like sum-calibrated ^{14}C data, if we leave aside the discussion of other methodological problems at this point. On the other hand, the sample size seems to have less influence on improving detection at some stage. Thus the systematic application of the simulation experiment of Contreras and Meadows (2014) cannot

Table 1: Results from the simulation (200 runs for each number of samples) of the original setup of (Contreras and Meadows 2014).

number of samples	sample density (per years)	mean proportion signal detected	standard deviation proportion signal detected	inner quartiles	95% quantiles
30	0.043	0.611	0.035	0.585–0.635	0.545–0.68
40	0.057	0.622	0.035	0.595–0.645	0.555–0.69
50	0.071	0.630	0.033	0.61–0.65	0.57–0.69
60	0.086	0.641	0.037	0.615–0.666	0.575–0.71
70	0.100	0.643	0.033	0.62–0.665	0.58–0.705
80	0.114	0.645	0.037	0.62–0.666	0.57–0.72
90	0.129	0.656	0.037	0.635–0.68	0.585–0.73
100	0.143	0.659	0.034	0.63–0.685	0.6–0.72
200	0.286	0.689	0.035	0.665–0.71	0.63–0.76
300	0.429	0.711	0.030	0.69–0.735	0.655–0.765
400	0.571	0.719	0.032	0.695–0.74	0.655–0.78
500	0.714	0.725	0.029	0.71–0.74	0.67–0.78
600	0.857	0.723	0.032	0.7–0.745	0.66–0.785
700	1.000	0.720	0.032	0.7–0.74	0.655–0.775
800	1.143	0.722	0.033	0.695–0.745	0.665–0.78
900	1.286	0.728	0.030	0.705–0.75	0.675–0.78
1000	1.429	0.721	0.029	0.7–0.74	0.665–0.77
2000	2.857	0.714	0.034	0.69–0.735	0.64–0.77

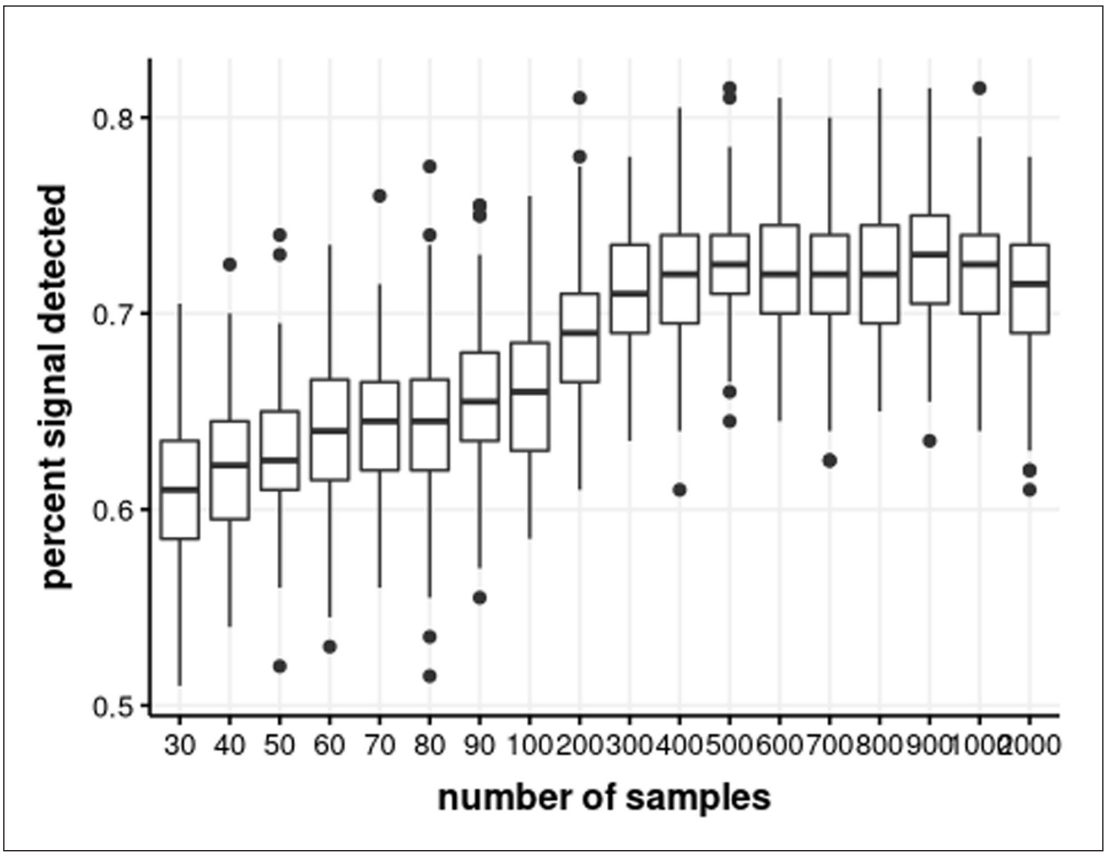

Figure 6: The results of the simulation of the original setup with 200 runs for each number of samples, visualised as boxplot (cf. Table 1).

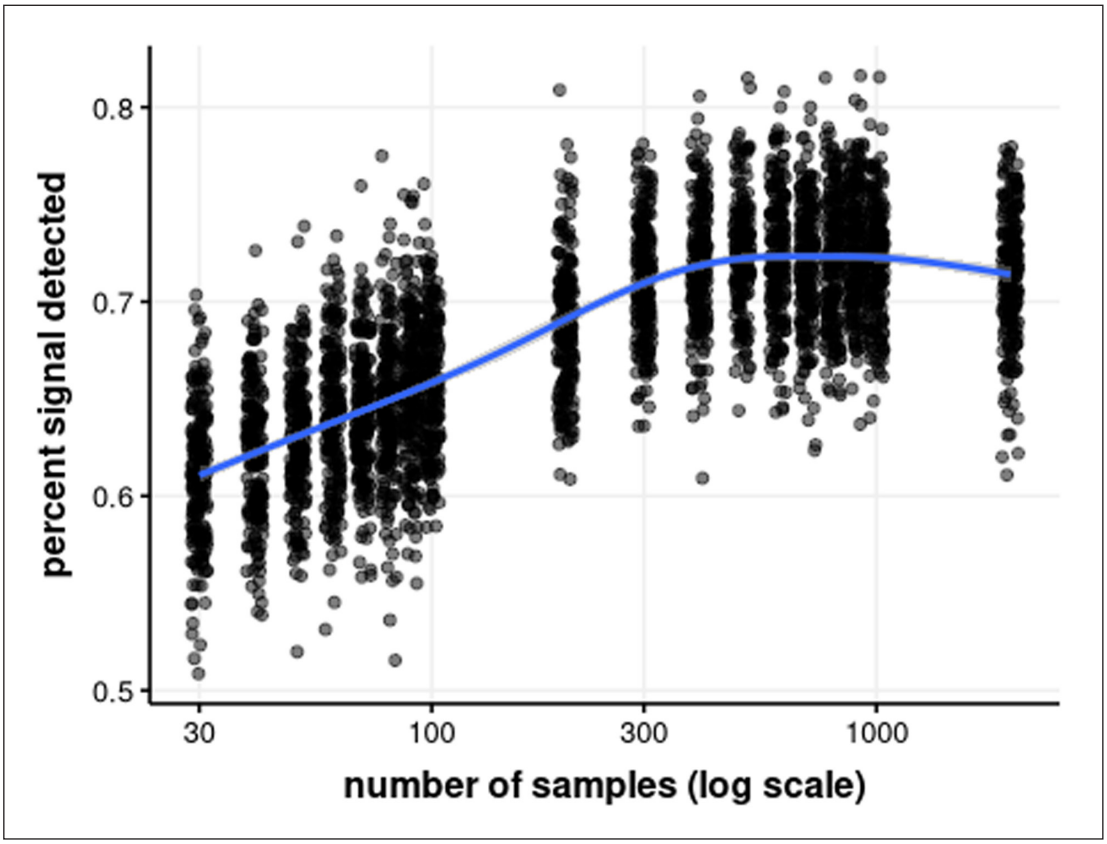

Figure 7: The results of the simulation of the original setup with 200 runs for each number of samples, visualised as plot with smoothed trend line (cf. Table 1). Note that the x-values are slightly jittered for better recognision of the individual dates, and the x-axis is logarithmic.

confirm the interpretations that they themselves deduce from their results. In this setup, it is not the number of samples that leads to a significant improvement in the detection rate. It is true that the individual results of individual sum calibrations deviate significantly from the given curve of the underlying population. But through formalized detection with fixed parameters, it is still possible to detect events within the given time window with a relatively high probability. Before we turn to the question of what exactly these events represent and how well we can separate false positive from true positive results (section 4.3), the influence of signal strength should be examined.

Altering Signal Strength

In the second part of the analysis, the intensity of the signal, as described above, was parameterised differently in order to check the influence of a stronger or weaker signal and thus be able to predict the detection possibilities of demographic changes of different intensity. **Figure 7** shows the mean detection rates for different scenarios. The signal strength originally used of 77.81 corresponds most closely to 0.8 in this parameterisation, which in this simulation leads to an average detection rate of 0.685 for 200 data or 0.73 for 1000 data. The results are therefore generally comparable with the reconstruction of the original simulation.

It is obvious that the strength of the signal has a high influence on the detection rate (**Figure 8**). Signals resulting from an underlying population reduced to 70% or less have a significantly higher detection rate, especially with higher sample numbers.

If the relationship between detection rate, sample size and signal strength is considered a linear model (see Appendix A.1.), then both factors are significant predictors for the detection rate. Signal strength (coefficient of 8.89e-05 with a p-value of 3.56e-08) is clearly more dominant than the sample size (coefficient of -6.53e-01 with a p-value of 3.78e-35).

It can be seen that a signal strength of 90% (corresponds to a reduction of 10%) with a small number of samples also shows a detection rate of more than 50%. This is rather surprising since the minimum difference necessary for recognition in the detection algorithm is set to 0.1. It is also surprising that this detection rate drops significantly with larger sample sizes (**Figure 8**). This is a strong indication that false-positive signals, which result exclusively from the random distribution of the data and not from the underlying pattern, are also counted here. This touches one of the key questions posed by Contreras and Meadows (2014): is it possible to distinguish real signals from false positives? To evaluate this, in a third step the same analysis was performed with the inclusion of a confidence envelope for false-positive signals.

Results with Testing for False Positives

In the same manner like the results above, **Table 3** and **Figure 9** visualise the effect of removal of false-positive pattern (section 3.4). In this version, the results for weak signals remain at a low level, while those for strong signals

Table 2: Results from the simulation of different signal strengths.

number of samples	sample density	0.3	0.4	0.5	0.6	0.7	0.8	0.9
30	0.043	0.805	0.760	0.680	0.680	0.700	0.550	0.605
40	0.057	0.805	0.775	0.700	0.680	0.640	0.610	0.615
50	0.071	0.860	0.860	0.720	0.775	0.660	0.605	0.600
60	0.086	0.905	0.820	0.745	0.705	0.645	0.635	0.510
70	0.100	0.880	0.820	0.790	0.750	0.695	0.625	0.545
80	0.114	0.855	0.815	0.785	0.735	0.675	0.645	0.595
90	0.129	0.850	0.855	0.820	0.720	0.650	0.675	0.585
100	0.143	0.950	0.885	0.835	0.755	0.685	0.640	0.610
200	0.286	0.980	0.950	0.880	0.845	0.730	0.685	0.620
300	0.429	1.000	0.990	0.920	0.885	0.840	0.680	0.565
400	0.571	1.000	0.990	0.975	0.900	0.820	0.645	0.595
500	0.714	1.000	1.000	0.985	0.900	0.835	0.680	0.535
600	0.857	1.000	1.000	0.975	0.945	0.795	0.635	0.505
700	1.000	1.000	1.000	0.990	0.935	0.845	0.705	0.510
800	1.143	1.000	1.000	0.990	0.955	0.810	0.685	0.505
900	1.286	1.000	1.000	0.990	0.970	0.850	0.730	0.485
1000	1.429	1.000	1.000	0.995	0.975	0.910	0.730	0.435
2000	2.857	1.000	1.000	1.000	0.985	0.885	0.615	0.365

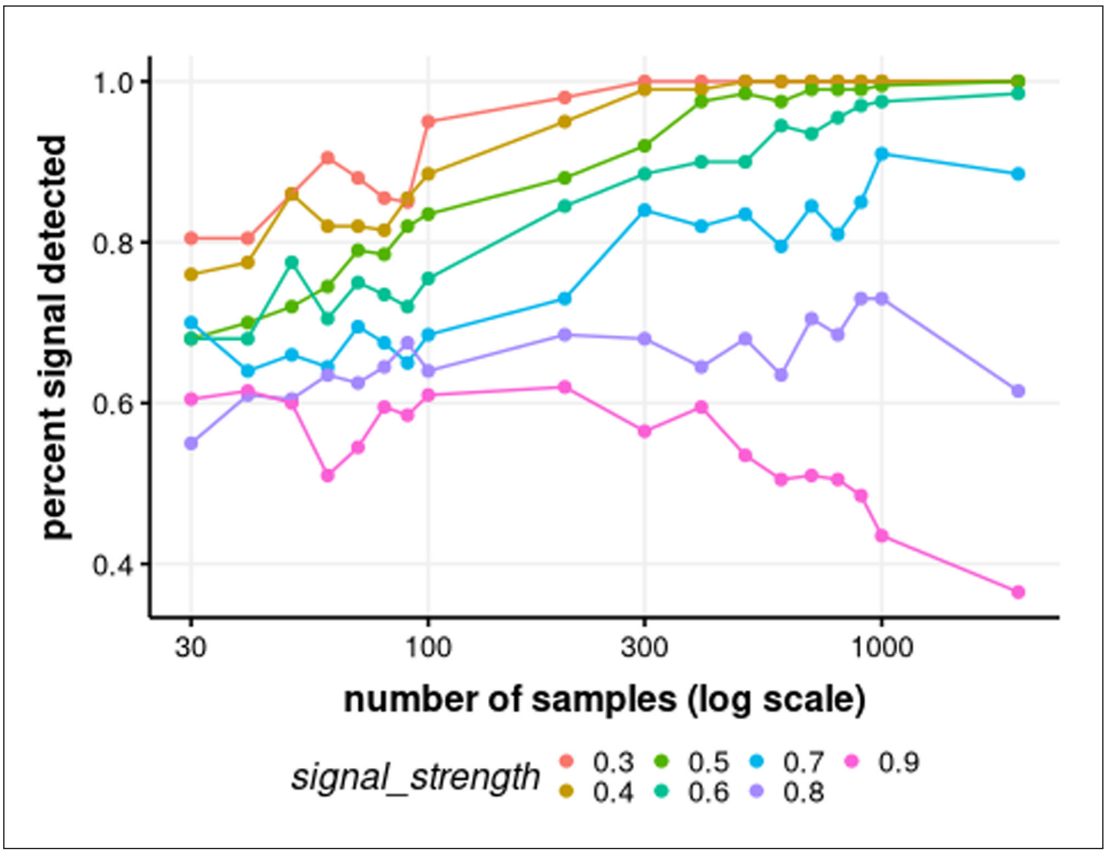

Figure 8: The results of the simulation of different signal strengths with 100 runs for each number of samples (cf. **Table 2**). Note that the x-axis is logarithmic.

Table 3: Results from the simulation of different signal strengths under consideration of the removal of false positive results.

number of samples	sample density	0.3	0.4	0.5	0.6	0.7	0.8	0.9
30	0.043	0.245	0.230	0.175	0.130	0.155	0.185	0.130
40	0.057	0.225	0.255	0.165	0.135	0.190	0.110	0.105
50	0.071	0.280	0.285	0.230	0.185	0.130	0.140	0.125
60	0.086	0.300	0.245	0.195	0.140	0.185	0.150	0.110
70	0.100	0.425	0.270	0.225	0.230	0.200	0.140	0.090
80	0.114	0.425	0.330	0.210	0.260	0.130	0.115	0.135
90	0.129	0.415	0.350	0.185	0.200	0.140	0.145	0.100
100	0.143	0.560	0.390	0.275	0.195	0.170	0.145	0.090
200	0.286	0.810	0.590	0.495	0.350	0.215	0.175	0.120
300	0.429	0.915	0.735	0.610	0.390	0.255	0.140	0.095
400	0.571	0.970	0.900	0.770	0.480	0.325	0.230	0.125
500	0.714	0.980	0.950	0.830	0.635	0.310	0.280	0.125
600	0.857	0.995	0.980	0.855	0.625	0.380	0.260	0.140
700	1.000	1.000	0.985	0.895	0.680	0.525	0.240	0.105
800	1.143	1.000	0.985	0.970	0.770	0.495	0.285	0.185
900	1.286	1.000	1.000	0.955	0.820	0.550	0.270	0.160
1000	1.429	1.000	1.000	0.960	0.855	0.625	0.360	0.110
2000	2.857	1.000	1.000	1.000	0.970	0.695	0.380	0.110

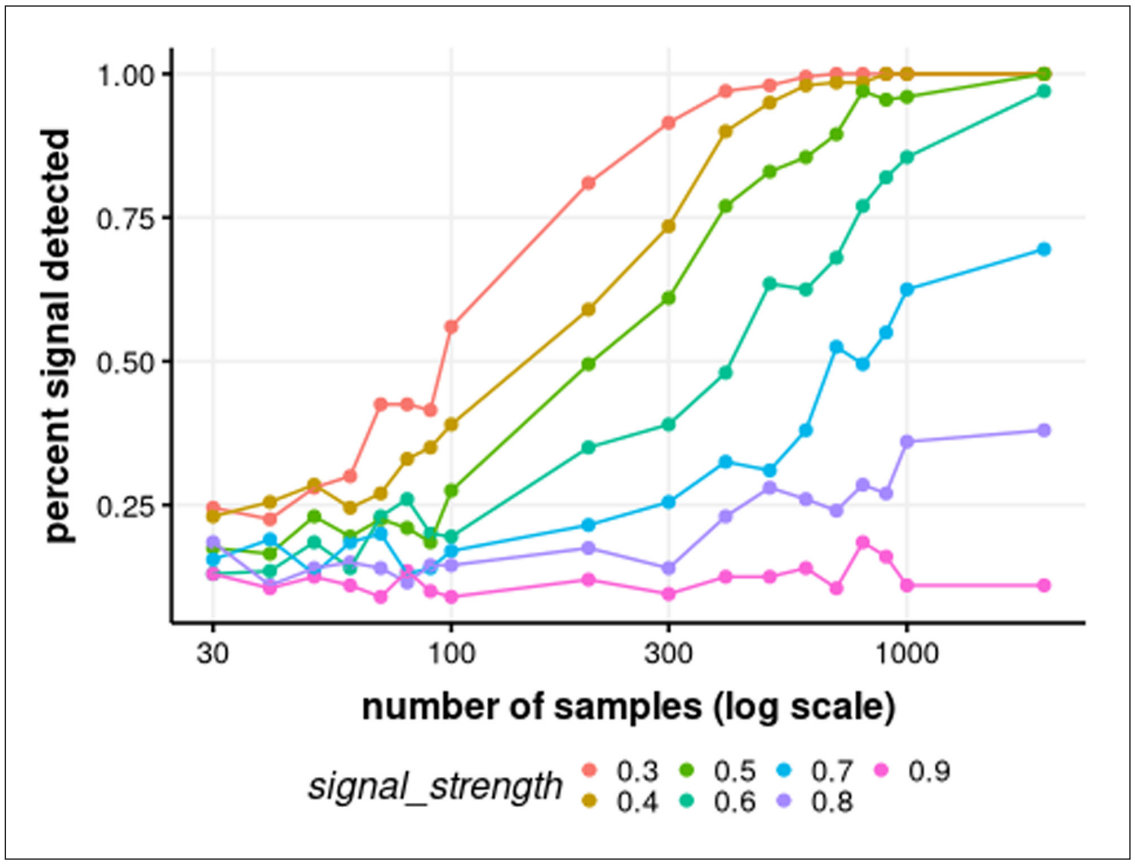

Figure 9: The results of the simulation of different signal strengths with 100 runs for each number of samples under consideration of the removal of false positive results (cf. Table 3). Note that the x-axis is logarithmic.

rise sharply from a sample size of about 200. For all signal strengths greater than 0.6, at the latest for a sample size of 300 or more, these exceed the 50% mark. This implies that this method produces a much more reliable result and is a strong indicator of the effectiveness of this approach. The overall detection rate is significantly reduced, and it becomes clear that for reliable identification of events a much higher sample size is necessary than if the possible false positives are naively ignored.

Finally, this combined method was applied again to the original example of the Black Death with its fixed signal strength. **Table 4** and **Figure 10** show the corresponding results. The scope of the sample size was extended upwards. Considering possible false-positive results, the detection rate for this pattern is quite low. For the scenario with 200 data points, corresponding to a density of 0.29 per year, there is a detection rate of 0.21, for a sample size of 1000, corresponding to a density of 1.43 per year, the detection rate reaches 0.325. Only with a sample size of 2000, (density = 2.86), a more or less reliable identification can be assumed.

Discussion and Conclusion

When using estimators for reconstructions, it is clear that in addition to the existing uncertainties, the variability in the relationship of the estimator to the estimated variable has to be considered. Therefore, it is unrealistic to expect a perfect reconstruction. Nevertheless, the question of Contreras and Meadows (2014) is, of course, justified as to whether such a disruptive event as the Black Death could have been recognised through this methodology. Therefore, the example is very well chosen and was used here for the same reason. The answer to this question must be 'yes', even if one has to limit, 'not in every case', or more precisely, 'with 68.9% probability', given the original setup of their analysis.

The other question raised by that paper is to what extent false positives can be distinguished from real signals. Here the approach of a hypothesis test based on bootstrapping was applied. This proved to be very effective to filter out false positive signals due to their lower magnitude. On the other hand, when this parameterisation is applied, an average of 5% remains which is not recognized as false positive. However, this is certainly an order of magnitude with which a science such as archaeology can operate, since very few approaches in our discipline offer 95 per cent confidence.

The detection of the event cannot be assumed to be absolutely certain. To the question of false-positive detections, the method of producing a confidence interval by simulation of an equal distribution (e.g. Shennan et al. 2013; Hinz et al. 2019) has been introduced into this simulation, going beyond the original setup of Contreras and Meadows. So if the question is how well we can identify this event, taking random fluctuations into account, the result is much more unfavourable. Only at a relatively high temporal density of ^{14}C dates a reliable detection becomes realistic.

A false-negative result, or in other words a Type II error, might be considered less dramatic in the given situation

Table 4: Results from the simulation of the original setup of Contreras and Meadows (2014) under consideration of the removal of false positive results.

number of samples	sample density	detection rate
30	0.043	0.145
40	0.057	0.165
50	0.071	0.145
60	0.086	0.120
70	0.100	0.100
80	0.114	0.130
90	0.129	0.140
100	0.143	0.115
200	0.286	0.210
300	0.429	0.160
400	0.571	0.235
500	0.714	0.285
600	0.857	0.170
700	1.000	0.320
800	1.143	0.365
900	1.286	0.330
1000	1.429	0.325
2000	2.857	0.395
3000	4.286	0.720
4000	5.714	0.710
5000	7.143	0.715
6000	8.571	0.705

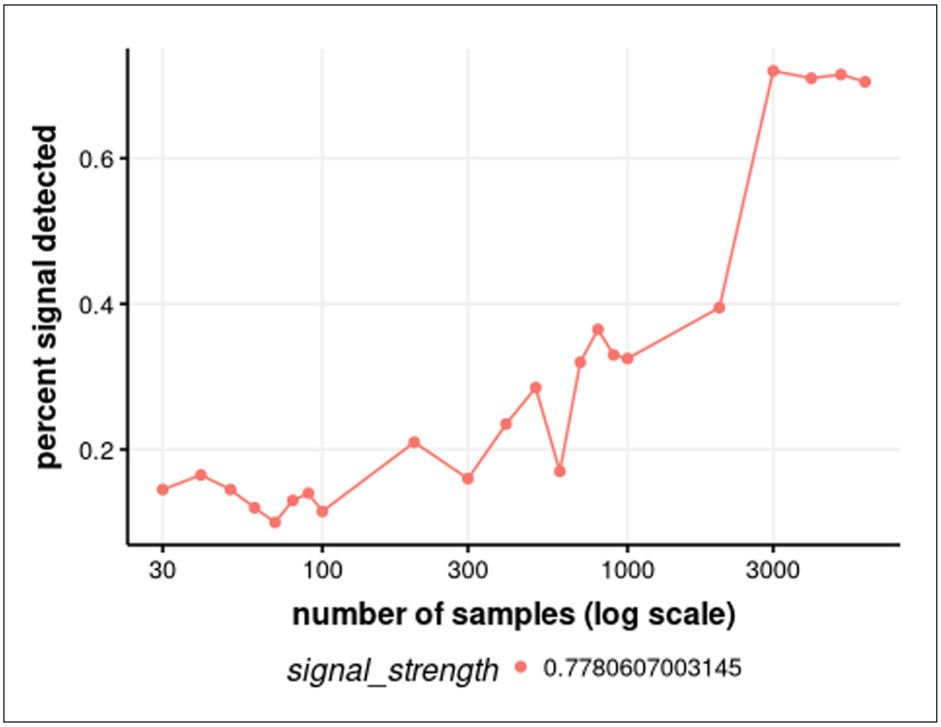

Figure 10: Results from the simulation of the original setup of Contreras and Meadows (2014) under consideration of the removal of false positive results (cf. tab. 4). Note that the x-axis is logarithmic.

than a Type I error where an event would be identified that is non-existent. There is some serious discussion as to which type I error would be worse and if there are any worse errors at all – this affects situations in which a type II error would lead to wrong decisions regarding, e.g., the safety of patient treatment (see, e.g,. Carlson 2017: 169–170). In this specific case, the methodology opens up the chance of detecting an event at the risk of not detecting it. In this instance, one should not assume that it may have been an uneventful period. This demands above all the necessity to not only reconstruct the past with one estimator but to validate different indicators mutually and to evaluate them in multi-proxy approaches.

The results raise questions about the nature of the conclusions already made using ^{14}C sum calibration. These results demonstrate distributions of ^{14}C data on the temporal axis that are different from the results of random sampling processes. The simulation results in this study clearly support the assumption that the significance test based on a Monte Carlo simulation, in one form or another, is very well suited to filter out signals that only appear in the data due to sampling effects. Therefore, significant results are highly likely to show variations in the data background. The basic question is therefore not of a statistical nature, and lies not in the insensitivity of the estimator itself, but rather in the fundamental methodological question of what information the absence of archaeological material at a certain period can provide us with. If, as in the case of the analysis of Shennan et al. (2013), we see a reduction of 36% (signal strength of 0.64) and take it literally (estimated by Shennan et al. (2013: 4), on a 200 years rolling mean, with a peak at about 3500 BCE and a minimum at about 3000 BCE), is this a population change that we can assume to be realistic for a prehistoric population? If we take the estimates of Müller (2015) into account, we are talking about 5 million people in Europe during this period. These would be reduced to about 3.2 million over 500 years. Another question is what true signal strength is indicated by an observed signal strength of 0.64, and how high the uncertainty range of such an estimate is. The latter estimate goes beyond the scope of this study but is an excellent question for further simulation-based studies.

The main problem in using summation calibration as a means of demographic reconstruction is therefore not the statistical conditions. If the sample size and sensitivity are too small, there are possibilities in this domain to identify such problems and to counteract them if necessary. The main problem lies rather in the often biased distribution caused by the production of the data in studies that almost never serve to produce a representative cross-section of dating in relation to the amount of material remains, but are usually carried out with specific different scientific objectives, as well as in the fact that often quite different deposition processes are treated equally. This is the original approach of Rick (1987), in which the amount of data was largely equated directly with the intensity of human activity, even when he already identified these biases. Therefore it is necessary to find appropriate countermeasures and to establish a best-practice catalogue for such investigations. An assessment of how strong a signal can be detected with what data density can be a valuable first step in the direction of such a standardisation.

In the course of the review of this paper, both reviewers independently suggested that the simulation should be performed for other temporal positions in order to check whether the results of the signal detection are robust to artifacts in the calibration curve. I do not see per se any methodological reasons why this would lead to significantly different results, since the curve in this period is quite comparable e.g. with that of the later Neolithic (a plateau between 1100–1200 CE and a wiggle between 1300–1400 CE, comparable e.g. with a wiggle between 3500–3400 BCE and a plateau between 3300–3100 BCE). Nevertheless, this is an interesting starting point for a possible further paper, but would go beyond the scope of the analysis presented here.

In this article a simulation approach was used to move beyond the simple statement 'the black plague could have remained undiscovered by ^{14}C sum calibration' and to arrive at a quantification of the probability of detection and a prediction of the detection potential of other, more or less pronounced events. As a result, it could be shown that no guarantee can be given for detection by this method, but that the chances will outweigh the risks.

Additional Files

The additional files for this article can be found as follows:

- **A.1.** Linear Model of detection rate, sample size and signal strength. DOI: https://doi.org/10.5334/jcaa.53.s1
- **A.2.** Colophon. DOI: https://doi.org/10.5334/jcaa.53.s2

Competing Interests

The author has no competing interests to declare.

References

Armit, I, Swindles, GT and **Becker, K.** 2013. From dates to demography in later prehistoric Ireland? Experimental approaches to the meta-analysis of large ^{14}C data-sets. *Journal of Archaeological Science*, 40(1): 433–438. DOI: https://doi.org/10.1016/j.jas.2012.08.039

Ballenger, JAM and **Mabry, JB.** 2011. Temporal frequency distributions of alluvium in the American Southwest: Taphonomic, paleohydraulic, and demographic implications. *Journal of Archaeological Science*, 38(6): 1314–1325. DOI: https://doi.org/10.1016/j.jas.2011.01.007

Bamforth, DB and **Grund, B.** 2012. Radiocarbon calibration curves, summed probability distributions, and early Paleoindian population trends in North America. *Journal of Archaeological Science*, 39(6): 1768–1774. DOI: https://doi.org/10.1016/j.jas.2012.01.017

Bayliss, A, Bronk Ramsey, C, van der Plicht, J and **Whittle, A.** 2007. Bradshaw and Bayes: Towards a Timetable for the Neolithic. *Cambridge*

Archaeological Journal, 17(Supplement S1): 1–28. DOI: https://doi.org/10.1017/S0959774307000145

Bleicher, N. 2013. Summed radiocarbon probability density functions cannot prove solar forcing of Central European lake-level changes. *The Holocene*, 23(5): 755–765. DOI: https://doi.org/10.1177/0959683612467478

Buchanan, B, Collard, M and **Edinborough, K.** 2008. Paleoindian demography and the extraterrestrial impact hypothesis. *Proceedings of the National Academy of Sciences*, 105(33): 11651–11654. DOI: https://doi.org/10.1073/pnas.0803762105

Carlson, DL. 2017. *Quantitative Methods in Archaeology Using R*. Cambridge Manuals in Archaeology. Cambridge University Press. DOI: https://doi.org/10.1017/9781139628730

Chiverrell, RC, Thorndycraft, VR and **Hoffmann, TO.** 2011. Cumulative probability functions and their role in evaluating the chronology of geomorphological events during the Holocene. *Journal of Quaternary Science*, 26(1): 76–85. DOI: https://doi.org/10.1002/jqs.1428

Collard, M, Ruttle, A, Buchanan, B and **O'Brien, MJ.** 2013. Population Size and Cultural Evolution in Nonindustrial Food-Producing Societies. *PLOS ONE*, 8(9): e72628. DOI: https://doi.org/10.1371/journal.pone.0072628

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

Crombé, P and **Robinson, E.** 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. *Journal of Archaeological Science*, 52: 558–566. DOI: https://doi.org/10.1016/j.jas.2014.02.001

Culleton, BJ. 2008. Crude demographic proxy reveals nothing about Paleoindian population. *Proceedings of the National Academy of Sciences*, 105(50): E111–E111. DOI: https://doi.org/10.1073/pnas.0809092106

Edinborough, K, Porčić, M, Martindale, A, Brown, TJ, Supernant, K and **Ames, KM.** 2017. Radiocarbon test for demographic events in written and oral history. *Proceedings of the National Academy of Sciences*, 114(47): 12436–12441. DOI: https://doi.org/10.1073/pnas.1713012114

Gamble, C, Davies, W, Pettitt, P, Hazelwood, L and **Richards, M.** 2005. The Archaeological and Genetic Foundations of the European Population during the Late Glacial: Implications for 'Agricultural Thinking'. *Cambridge Archaeological Journal*, 15(02): 193–223. DOI: https://doi.org/10.1017/S0959774305000107

Gkiasta, M, Russell, T, Shennan, S and **Steele, J.** 2003. Neolithic transition in Europe: The radiocarbon record revisited. *Antiquity*, 77(295): 45–62. DOI: https://doi.org/10.1017/S0003598X00061330

Hinz, M, Feeser, I, Sjögren, K-G and **Müller, J.** 2012. Demography and the intensity of cultural activities: An evaluation of Funnel Beaker Societies (4200–2800 cal BC). *Journal of Archaeological Science*, 39(10): 3331–3340. DOI: https://doi.org/10.1016/j.jas.2012.05.028

Hinz, M, Schirrmacher, J, Kneisel, J, Rinne, C and **Weinelt, M.** 2019. The Chalcolithic–Bronze Age transition in southern Iberia under the influence of the 4.2 kyr event? A correlation of climatological and demographic proxies. *Journal of Neolithic Archaeology*, 21: 1–26–1–26. DOI: https://doi.org/10.12766/jna.2019.1

Hinz, M, Schmid, C, Knitter, D and **Tietze, C.** 2018. *oxcAAR: Interface to 'OxCal' Radiocarbon Calibration*. Available at https://rdrr.io/cran/oxcAAR/.

Hoffmann, T, Lang, A and **Dikau, R.** 2008. Holocene river activity: Analysing ^{14}C-dated fluvial and colluvial sediments from Germany. *Quaternary Science Reviews*, 27(21–22): 2031–2040. DOI: https://doi.org/10.1016/j.quascirev.2008.06.014

Johnstone, E, Macklin, MG and **Lewin, J.** 2006. The development and application of a database of radiocarbon-dated Holocene fluvial deposits in Great Britain. *CATENA*, 66(1–2): 14–23. DOI: https://doi.org/10.1016/j.catena.2005.07.006

Kelly, RL, Surovell, TA, Shuman, BN and **Smith, GM.** 2013. A continuous climatic impact on Holocene human population in the Rocky Mountains. *Proceedings of the National Academy of Sciences*, 110(2): 443–447. DOI: https://doi.org/10.1073/pnas.1201341110

Marwick, B. 2017. Computational Reproducibility in Archaeological Research: Basic Principles and a Case Study of Their Implementation. *Journal of Archaeological Method and Theory*, 24(2): 424–450. DOI: https://doi.org/10.1007/s10816-015-9272-9

McLaughlin, TR. 2019. On Applications of Space–Time Modelling with Open-Source ^{14}C Age Calibration. *Journal of Archaeological Method and Theory*, 26(2): 479–501. DOI: https://doi.org/10.1007/s10816-018-9381-3

Müller, J. 2015. 8 Million Neolithic Europeans: Social Demography and Social Archaeology on the Scope of Change – from the Near East to Scandinavia. In: Neustupný, E and Kristiansen, K (eds.), *Paradigm found: Archaeological theory – past, present and future: Essays in honour of Evžen Neustupný*, 200–214. Oxford: Oxbow Books. DOI: https://doi.org/10.2307/j.ctvh1dpc1.20

Mulrooney, MA. 2013. An island-wide assessment of the chronology of settlement and land use on Rapa Nui (Easter Island) based on radiocarbon data. *Journal of Archaeological Science*, 40(12): 4377–4399. DOI: https://doi.org/10.1016/j.jas.2013.06.020

Prates, L, Politis, G and **Steele, J.** 2013. Radiocarbon chronology of the early human occupation of

Argentina. *Quaternary International*, 301: 104–122. DOI: https://doi.org/10.1016/j.quaint.2013.03.011

Rick, JW. 1987. Dates as Data: An Examination of the Peruvian Preceramic Radiocarbon Record. *American Antiquity*, 52(1): 55–73. DOI: https://doi.org/10.2307/281060

Riede, F. 2009. Climate and Demography in Early Prehistory: Using Calibrated ^{14}C Dates as Population Proxies. *Human Biology*, 81(3): 309–338. DOI: https://doi.org/10.3378/027.081.0311

Rieth, TM, Hunt, TL, Lipo, C and **Wilmshurst, JM.** 2011. The 13th century polynesian colonization of Hawai'i Island. *Journal of Archaeological Science*, 38(10): 2740–2749. DOI: https://doi.org/10.1016/j.jas.2011.06.017

Shennan, S. 2009. Evolutionary Demography and the Population History of the European Early Neolithic. *Human Biology*, 81(2–3): 339–355. DOI: https://doi.org/10.3378/027.081.0312

Shennan, S. 2012. Demographic Continuities and Discontinuities in Neolithic Europe: Evidence, Methods and Implications. *Journal of Archaeological Method and Theory*, 20(2): 300–311. DOI: https://doi.org/10.1007/s10816-012-9154-3

Shennan, S, Downey, SS, Timpson, A, Edinborough, K, Colledge, S, Kerig, T, Manning, K and **Thomas, MG.** 2013. Regional population collapse followed initial agriculture booms in mid-Holocene Europe. *Nature Communications*, 4: 2486. DOI: https://doi.org/10.1038/ncomms3486

Shennan, S and **Edinborough, K.** 2007. Prehistoric population history: From the Late Glacial to the Late Neolithic in Central and Northern Europe. *Journal of Archaeological Science*, 34(8): 1339–1345. DOI: https://doi.org/10.1016/j.jas.2006.10.031

Steele, J. 2010. Radiocarbon dates as data: Quantitative strategies for estimating colonization front speeds and event densities. *Journal of Archaeological Science*, 37(8): 2017–2030. DOI: https://doi.org/10.1016/j.jas.2010.03.007

Surovell, TA and **Brantingham, PJ.** 2007. A note on the use of temporal frequency distributions in studies of prehistoric demography. *Journal of Archaeological Science*, 34(11): 1868–1877. DOI: https://doi.org/10.1016/j.jas.2007.01.003

Surovell, TA, Byrd Finley, J, Smith, GM, Brantingham, PJ and **Kelly, R.** 2009. Correcting temporal frequency distributions for taphonomic bias. *Journal of Archaeological Science*, 36(8): 1715–1724. DOI: https://doi.org/10.1016/j.jas.2009.03.029

Tallavaara, M, Pesonen, P and **Oinonen, M.** 2010. Prehistoric population history in eastern Fennoscandia. *Journal of Archaeological Science*, 37(2): 251–260. DOI: https://doi.org/10.1016/j.jas.2009.09.035

Timpson, A, Colledge, S, Crema, E, Edinborough, K, Kerig, T, Manning, K, Thomas, MG and **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. DOI: https://doi.org/10.1016/j.jas.2014.08.011

Torfing, T. 2015. Neolithic population and summed probability distribution of ^{14}C-dates. *Journal of Archaeological Science*. DOI: https://doi.org/10.1016/j.jas.2015.06.004

Whitehouse, NJ, Schulting, RJ, McClatchie, M, Barratt, P, McLaughlin, TR, Bogaard, A, Colledge, S, Marchant, R, Gaffrey, J and **Bunting, MJ.** 2014. Neolithic agriculture on the European western frontier: The boom and bust of early farming in Ireland. *Journal of Archaeological Science*, 51: 181–205. DOI: https://doi.org/10.1016/j.jas.2013.08.009

Williams, AN. 2012. The use of summed radiocarbon probability distributions in archaeology: A review of methods. *Journal of Archaeological Science*, 39(3): 578–589. DOI: https://doi.org/10.1016/j.jas.2011.07.014

How to cite this article: Hinz, M. 2020. Sensitivity of Radiocarbon Sum Calibration. *Journal of Computer Applications in Archaeology,* 3(1): 238–252. DOI: https://doi.org/10.5334/jcaa.53

Submitted: 20 January 2020 **Accepted:** 26 June 2020 **Published:** 12 August 2020