Reviews Actionable Items

Martin Hinz

16 6 2020

Reviewer A

- ☒ It probably should be noted that McLaughlin 2019 undertook a similar simulation experiment with the Contreras and Meadows black death case study, using KDEs of radiocarbon (with a 94% detection rate). The influence of sample size and the risk of false positives were not however addressed, so this new paper can be seen as significant progress
  - Many thanks for the appreciation! The fact that this paper was not cited is an obvious omission which has been corrected in the revised version.
- ☒ An useful analytical step not undertaken in the current paper (but was one of the issues discussed by Contreras and Meadows) would be the transposition of the black death signal to a different region of time, to check that the signal detection results are robust against artifacts in the calibration curve.
  - I added a footnote commenting on this, also in respect to the remark of Reviewer D
- ☒ Writing is excellent; there are a couple of typos L239-240 'It have to be noted' and L255 'ab auctoritate'.
  - Corrected.
- ☒ The author should probably comment on Edinborough et al 2017 PNAS hypothesis test for short term events in archaeological radiocarbon
  - A very valuable advice. The paper mentioned above takes a different approach and, in my opinion, requires a certain amount of prior knowledge in order to choose the right points in time for such a test. Since this is not available for the scenarios on which this paper is focused, especially those of prehistoric times, the approach was not pursued further in this paper. Nevertheless, I have added a paragraph to acknowledge the contribution.

Reviewer D

Review of "Sensitivity of Radiocarbon Sum Calibration" by Hinz

- ☒ They are limited in that only one, fairly short time-scale is explored. The author might comment on the applicability of this to, for example, detecting a population crash in the late Neolithic.
  - I added a footnote commenting on this, also in respect to the remark of Reviewer A
- ☒ My fundamental issue with this paper is that it does not recognise that estimating the detection rate is fundamentally estimating a probability from a binomial count. Lines 151-153 indicate rather vaguely that 200 batches of 200 simulations were performed; I only understood this by looking at the code. In Table 1 the standard deviation, inner quartiles and 95% quantiles are simply dependent on the size of the batches (200) and underlying true detection rate. The standard deviation of a binomial estimate expressed as percentage is 100*sqrt(pq/(n+1)). For n=200 and values of p from 61.1% to 72.8%, the sd ranges 3.44% to 3.14%, very close to what is reported empirically, especially as the latter will also

include some noise as it is simulated. 200 batches of 200 simulations doesn't tell us anything more than 40000 simulations on their own. For a given scenario, columns 4-6 of Table 1 will vary with batch size in a way that is predictable once the success rate is known. With n=40000 the s.d. is about 0.23% to 0.24% here. For the later simulations with n=100, the s.d. will not be more than 5%. Consequently parts of the discussion (lines 273-4, 276-7) and presentation (sizes of boxes and whiskers in Fig.6, scatter of points in Fig.7) are not relevant as these are simply functions of the chosen batch size. I am grateful for this thoughtful commentary, which suggests a deep statistical understanding on the part of the reviewer. The procedure was chosen because the non-linear nature of the calibration curve could not exclude a non-linear behaviour of the standard deviations. The primary objective is to estimate for a single sum calibration how reliable the results are and which range of variation (sd) can be expected. The standard deviation is also partly dependent on the sample size of radiocarbon data, since a higher noise factor resulting from less data also increases the volatility in detection, as is shown in the theory of sampling and as can be empirically deduced from the figures here as well. I therefore consider it justified to present the results in the chosen form, as this allows the reader to assess the reliability of the detection rate estimate and to draw conclusions about the range of variation to be expected for individual sum calibrations. However, I was happy to extend the presentation of the methodology to include the elements mentioned by the reviewer.

☒ With se of the estimate at 0.2% only the first decimal place should be reported in all the estimated percentages, or third decimal place if using proportions. Table 1 uses percentages but Tables 2 and 3 use proportions – this should be made consistent and proportions would be better.

- I have implemented the proposals and changed the rounding for the percentages to 1 decimal place, while changed table 1 to proportions.

☒ Lines 345 and 346: this is not 'significance' (which is either there or not) but a p-value from which we decide on significance.

- That is absolutely correct, I have changed it accordingly.

Other comments

☒ Lines 148-154: what are scenarios in the plural here? The start of the paragraph says it is using the single scenario of Contreras & Meadows.

- Thank you for pointing this out! The different scenarios refer to different data densities, while the general scenario (the Black Plague) remains unchanged. I clarified that in the text.

☒ Lines 200-203: it would be useful to have these date ranges marked on Figures 4 and 5.

- Thanks for this advice! I have changed the illustrations and captions accordingly, the ranges are now shown in blue.

☒ Lines 204-205: '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' I don't understand what this means. Is it that, for example, if the minimum is at 1400 it is compared to the means of 1250-1350 and 1450-1550?

- Admittedly, it was not written very clearly. The event as such is assumed to be fixed, it has to be between 1260 and 1580 (the test range +- 50 years, respectively). From these dates in each case +- 50 years is taken as the reference period. It is now clarified accordingly in the text.

☒ Line 280 seems to be in error stating 'detection chance seems to be relatively independent of the sample size', as line 291-3 states 'Up to a sample size of about 300 ... the detection rate improves and then remains in a plateau' and this is clear in Fig.7

- That is indeed unclearly expressed, thank you very much for the note! Such low densities have also generally been regarded as inadequate up to now, so the wording chosen previously ignored the issue. That has now been changed in the revised version.

☒ In figures 6 and 7 why is there switch to 100 simulation rather than the 200 of Table 1?

- Thanks for observing this error, which resulted from an old parameterization (with 100 runs). This is corrected to 200 in the captions in the modified version.

☒ The code at lines 323-342 is not necessary in the main text and could go in an appendix

- A very good suggestion! I have created an appendix section in which the model was inserted as A.1 and the colophon as A.2.

☒ Line 394: 68.88% is not justified and no one is interested in knowing the success rate to this precision. 69% will do.

- This seems to make perfect sense to me, I have changed the rounding accordingly.

Minor corrections

☒ Figure 1 contains a typographic error: 'certanity' should be 'certainty'
☒ Line 364: replace 'resp.' with 'and'
☒ Table 4: what is the heading '0.7780607003145'?
☒ Line 379: 'The table 4 or Fig. 10' should be 'Table 4 and Fig. 10'
☒ Line 448: 'assessement' should be 'assessment'