Review of

The conflict between sampling resolution and stratigraphic constraints from a Bayesian perspective: OSL and radiocarbon case studies

28th June 2025

Overview

This paper presents two worked case study examples that show three different modelling approaches can lead to strikingly different results when applied to the same data.

I think this is an interesting result, and important for readers who might be using these models to understand, but I do feel that the current exposition is really quite unclear – in terms of sufficiently describing what the different models do; their interpretation; and the potential reasons for any differences in the resulting inference.

They argue that the differences they see between models are due to the way that they handle stratigraphic ordering, but it is not entirely clear to me that it is solely this – as they also seem to argue that the models implement this stratigraphic information in very similar ways.

If that is the case, then I would expect similar results across the models. Instead it suggests to me that either they actually handle stratigraphy quite differently, or that there are more fundamental differences between the three models (which are not explained) that effectively lead to quite different modelling assumptions; or that (some of) the models have perhaps not converged correctly.

Specifically, they compare:

1. BayLum when modelling OSL dates
2. OxCal, BayLum and ChronoModel when used to analyse a selection of 14C dates from Catalhoyuk

As I said, I think the overall manuscript has highly useful and valuable content for the community and I would recommend publication, but IMO the overall narrative and level of clear explanation really does need to be improved if it is to be of substantial use to the relevant community.

<u>Major Comments</u>

<u>Explanation of the Models</u>

There needs to be much more, and understandable, detail on the mechanics of the three actual models. Currently, there is quite a lot of written text (which I feel is hard to parse and repetitive in places) but I was not particularly clear on how any of the three approaches actually modelled the data – in particular BayLum or ChronoModel as these are less well-known to me.

I do not think that in this instance, where you are arguing the models are inconsistent with one another, it is sufficient just to refer to the original papers. One needs to understand why the models might be presenting different results to one another (and that must either be because they have different underlying statistical modelling approaches, or that the MCMC they all seem to use hasn't converged properly (which is a serious issue with using MCMC on an ordered parameter space as it becomes highly multimodal).

I understand you cannot describe the entire approach for each, but what are the actual mathematical/statistical models for each that are fitted (what is the likelihood, what is the prior, …). In my view this should be done through short mathematical equations with a suitable notation, not long passages of text which aren't sufficiently explicit.

If it is as simple as just some phase models and stratigraphy I would hope/assume that this can be provided in a couple of general equation (see e.g., Nicholls and Jones, 2001, for a complete and general explanation of the stratigraphic phase model as I presume is fitted in OxCal to Catalhoyuk).

Comparison with Nicholls and Jones (2001)

It is not clear to me that the authors have entirely understood this paper as they seem to argue against the point I see it as trying to make several times (e.g., line 51) while then citing it as an example of an approach to address the problem of *under* spreading as they see it over. Importantly, I do not see this paper as making a particular statement about using a uniform phase within a boundary. It is about a different aspect – the prior on phase boundaries.

My reading/understanding of Nicholls and Jones (2001) is that their main thrust is to argue that if you do not use a sensible prior then you get erroneous **overestimation** of the spread of stratigraphic calendar ages (i.e., dates that are too far spread). Hence they are arguing precisely that you should compress/shrink the calendar ages. This paper evidently needs to be referenced but I think in a very different way.

Nicholls and Jones' specific, recommended, prior actively aims to penalise the spread of phase boundaries. This problem of over-spreading is well known and theoretically evidenced (e.g., by Stein, 1956). Nicholls and Jones demonstrate this in the context of 14C calibration convincingly using Bayes factors. They also provide a discussion of what underlying depositional model leads to their chosen prior which penalises the spread, and an argument it is more sensible than the model which generates a constant prior density. They are therefore arguing against the idea of this paper (i.e., they are making a case things should be compressed!).

Nicholls and Jones (2001) consider a model that consists of ordered phase layers (with adjoining boundaries $\psi_0, \ldots, \psi_M$) and within each layer they assume there are a set of samples with calendar ages. They assume no ordering within the samples inside a layer. This seems to be precisely the model you fit to Catalhoyuk.

Specifically, they assume that the N_i ^{14}C samples within a specific layer/phase i (i.e., the samples between calendar ages ψ_{i-1} and ψ_i) have calendar age $\theta_{i,1}, \ldots, \theta_{i,N_i}$ and are uniformly distributed in calendar time through the layer (i.e., there is no information, or prior, placed, on their relative ordering). The ordering is really only placed on the phase boundaries, i.e., $\psi_i > \psi_{i-1}$

Throughout their paper, they propose a completely consistent prior on the individual sample calendar ages $\theta_{i,j}$ (conditional on the phase boundaries). This is a uniform prior because they don't assume any ordering of the samples within the layer.

Their argument is entirely concerned with what prior you should place on the phase boundaries. They propose two – a constant prior that does not penalise the overall range/span of all the layers which they argue is actually highly informative on the overall span and leads to overspread inference; and one that does, i.e., operate to reduce the overall span $\psi_M - \psi_0$ which they argue (with theoretical support) works better.

Which it the bit you are arguing against? The uniform phase component of the model within each layer? Or the prior on the spacings of the phase boundaries. These are quite different things.

As far as I am aware, Nicholls and Jones (2001) do not discuss, or argue for/against the uniform prior on the samples within each layer – this is taken as an assumption of both their priors/models.

If I am correct then I think this therefore means that some needs some significant rewordings are needed in the paper. There is stuff about uniform phase models being needed to reduce erroneous over-spread (where the samples do arise from such a model) but it I'm not sure this is Jones and Nicholls (2001).

Assessment of Model Appropriateness

I entirely agree with the authors that all users should not treat statistical models as black-boxes, but rather ensure/investigate that the assumptions made within those models are valid for their example/analysis. Also that there is no such thing as a non-informative prior (and that the choice of prior can have large consequences on inference). This is something which is far too frequently overlooked, especially in Bayesian analyses (which will always give you an answer even if it doesn't make sense).

However, I am a little concerned by the idea/conclusion which I feel goes too far in suggesting that you should perhaps ditch modelling altogether, and that you can conclude the models are wrong primarily because they don't overlap with selected independently-calibrated 14C dates. There is a large literature (backed up by theory) which tells you that independent estimation of any random variables is not a statistically-valid approach and leads to overly-spread estimation (going back to Stein, 1956). We should not encourage people to go back to that.

Of course, one should be concerned by the differences between the inference provided by these models but I think the reasons are more nuanced than you present. As you say none of the models are likely correct – but then neither is independent calibration of each sample. Presumably some of the models fundamentally have components which you subjectively might not support in their construction/prior. I feel you need to draw that out (which would be easier if the actual statistical modelling equations were explicitly laid out)

Technical/Minor Comments

Equation 2 seems wrong to me. Should the numerator on the RHS be 1; and the product moved outside of the fraction? Currently the RHS numerator repeats the LHS – which is circular.

BayLum uses IntCal13 – I would suggest that you need to show/state that this does not lead to the differences in inference. It shouldn't as IntCal13 and IntCal20 are very similar in the period of Catalhoyuk but you need to explain/show this (e.g., by providing plot of the two IntCal curves alongside one another in this interval).

I commend the authors on making their code available but it would be nice if it was on Zenodo/Github rather than simply text in an Appendix.