RADIOCARBON, Vol 42, Nr 2, 2000, p 199–202

COMMENT ON 'THE USE OF BAYESIAN STATISTICS FOR ^{14}C DATES OF CHRONOLOGICALLY ORDERED SAMPLES: A CRITICAL ANALYSIS'

Christopher Bronk Ramsey

University of Oxford Radiocarbon Accelerator Unit, 6 Keble Road, Oxford, OX1 3QJ, United Kingdom.
Email: christopher.ramsey@rlaha.ox.ac.uk

INTRODUCTION

In this issue, the paper "The Use of Bayesian Statistics for ^{14}C Dates of Chronologically Ordered Samples: A Critical Analysis" by Steier and Rom brings up some interesting points that help to focus discussion on this subject. There are, in fact, two distinct issues here:

- What prior should be used for groups of dated samples
- What the results of the Bayesian analysis actually mean

The first of these is interesting because, although considered carefully from the early use of Bayesian analysis for radiocarbon dates, it is still often overlooked in its application. The second question must be put in the context of the wider debate that rages in the ever-increasing number of scientific disciplines where Bayesian statistics are being used.

The Multiple Samples Problem

The focus of this paper is on the inappropriate nature of the prior outlined in equation (6) for a sequence of samples. What the authors do not notice is that this prior is also inappropriate if the samples are not constrained to be in order. We can look at this mathematically and then see why it makes sense intuitively.

Mathematically the problem derives from the fact that the number of possible combinations of dates with a given overall span is proportional to s^{n-2} where s is the span of the dates and n is the total number. In the case of samples that have a defined sequence, as discussed in this paper, the number of possible states is reduced by a factor of $n!$ but this does not alter the bias on the span. However, this bias does indeed explain the results of computer experiments A and B, as identified by Steier and Rom.

We can also understand what is going on here intuitively. The rationale behind the "neutral" prior is that the samples are randomly derived from a continuum of events that are Poisson distributed. The assumption is that these events are chosen from an infinite time slice and therefore *a priori* are very unlikely all to be of similar age; hence the strong bias towards a long span. In practice, any such selection of samples is derived from a finite time slice and it is at least reasonable to assume that they might be Poisson distributed from within that period.

Boundaries

A way around this problem was first described early in the use of Bayesian statistics for analysis of ^{14}C dates (Buck et al. 1992) and incorporated in the first release of OxCal (Bronk Ramsey 1995) and applied in many analyses (e.g. Buck et al. 1992; Bayliss et al. 1997; Needham et al. 1998).

Mathematically this method can be seen as applying a prior of $s^{-(n-2)}$ to each possible combination, as indicated in equation (16) of this paper or explained in Bronk Ramsey (1999). This is not, however, the exact implementation used by Buck et al. or OxCal. In these cases, two additional undated events are postulated: a start event and an end event that bracket the dates under consideration. In

OxCal these events are termed boundaries, so for a phase of multiple samples one would have (see Bronk Ramsey 1998 for notation):

```
Sequence { Boundary; Phase { ... }; Boundary; };
```

Or in the case of a sequence, as discussed in this paper:

```
Sequence { Boundary; Sequence { ... }; Boundary;};
```

Assuming again that there are n events within the main group there are now $n+2$ events in total, including the boundaries. The prior is therefore biased by s^{-n} to overcome the effect of the problems described above. When analyses are conducted like those described in computer experiments A and B, the results using this prior are in agreement with expectations and do not show the effects described.

Intuitively, what we are doing here is recognizing that the events under consideration are likely to be selected from a finite time slice that lies between the two undated boundaries. We define a prior for the interval between these two boundaries as being uniform (i.e. any length is equally likely on a linear scale). This is a neutral assumption on the overall span although one could argue that other distributions might be more reasonable in some circumstances. Between the two boundaries, the events are still assumed to be Poisson distributed.

In practice, most real archaeological sites or cultures can be split up into a number of different phases. In these cases, there can be several boundaries. This introduces further considerations, which are discussed in the manual for v3.3 of OxCal (see http://www.rlaha.ox.ac.uk). The prior defined in this way is not arbitrary or vague. It is based on a model that the events we sample are drawn from a Poisson distribution during periods of equilibria punctuated by boundaries, which are themselves Poisson distributed in time.

The Meaning of the Results of Bayesian Analysis

We have still not addressed computer experiment C, which relates to only two events. Although these could be bracketed by boundaries as described above, this would not overcome the effect described. Here we need to be clear what any Bayesian result means.

The age range quoted (at, for example 95%) from Bayesian analysis (including the widely used probability method of calibration) is a range of values that includes the 95% most likely results based on the assumed prior. It does *not* mean that we can be 95% confident that any result lying outside this range is false. This contrasts with the intercept method of calibration, which is based on a classical null-hypothesis. Likewise, it is to be expected that if we select some particular situation, Bayesian statistics may not give us the correct result 95% of the time. What it should do is give us results that are correct in 95% of possible situations.

It should be stressed that this will be the case for a single calibrated date too. Let us consider, for example, ^{14}C dating an object that dates from 405 BC which is just off the main 1st Millennium BC plateau. Furthermore, let the precision be only ±60 as this exacerbates the effect of the plateau. The calibrated ranges (at 95%) only include 405 BC about half as often if the probability method is used instead of the intercept method. Anyone can test this using R_Simulate in OxCal and the INTCAL98 calibration curve. This need not necessarily worry us as it is very hard to find a date for which this is the case and, for 95% of dates, the probability method will give us the correct answer.

This underscores the principal difference between Bayesian and classical statistics. If we have a definite hypothesis (like this sample is a particular age, or all of these samples are the same age), classical statistics provide us with the tools for testing this hypothesis. If, on the other hand, as is often the case in dating projects, our possible solutions are virtually infinite, Bayesian statistics provide us with a way to pick out the most likely possibilities. The priors define our set of possibilities and do therefore need choosing with care, as the authors of this paper demonstrate.

Choosing Priors

The choice of priors is always the hardest part of any Bayesian analysis. In practice, it often does not make as much difference as you would think. For example, if we ignore the effect shown in computer experiments A and B of this program, and do not use boundaries in our analysis, this has very little effect on many archaeological sites. This is because the dates are often reasonably well resolved and not very high in number. In effect, the boundaries themselves are often effectively provided by other archaeological information. This said, it is clearly best to use a realistic prior and anyone attempting this sort of analysis should certainly seek advice first from someone familiar with the technique. In some cases, the use of boundaries might not be appropriate.

There is no one correct prior for a given situation. A prior is merely a model that can be applied to the data to help in its interpretation. Ideally, several different models with different priors should be tried. If the results from these are all similar, it demonstrates that the conclusions are insensitive to the prior. This approach may not appeal to some people who wish to take their results, process them statistically and come out with the "right" answer. However, it is not really so very different from the hypothesis testing method of classical statistics. Here, instead of testing several hypotheses, we apply several prior models based on different possible interpretations of the data.

In addition, some checks can be applied to prior models. The agreement indices of OxCal, for example, will highlight cases where the prior model is inconsistent with the data. This is, in fact the case if computer experiments A and B are attempted with OxCal. This feature also allows us to test models against the data to see if they are consistent (as in Needham et al. 1998). It should not, however be relied upon to identify all inappropriate priors, as it will only do so if the measurement data themselves are inconsistent with the prior.

CONCLUSION

Steier and Rom's paper highlights one potential problem with choosing priors for Bayesian analysis. The authors identify this with sequences, although mathematically it is equally true for any group of events. The issue raised is one that has been considered in some detail for many years but does still need discussion since it is also often ignored in applications. The authors correctly suggest that those starting to undertake this kind of analysis should ask for advice.

Also raised here is the nature of Bayesian analysis and this provides a useful stimulus to consider the fundamental differences between classical and Bayesian statistics and the different applications for which they are best suited. Used properly, Bayesian analysis can be a powerful tool to help in the interpretation of radiocarbon dates and enable us to draw together various forms of information in a way that is simply not possible by using classical statistics alone.

REFERENCES

Bayliss A, Bronk Ramsey C, McCormac FG. 1997. Dating Stonehenge. In: Cunliffe B, Renfrew C, editors. *Science and Stonehenge*. Proceedings of the British Academy 92:39–59.

Bronk Ramsey C. 1995. Radiocarbon calibration and analysis of stratigraphy: the OxCal program. *Radiocarbon* 37(2):425–30.

Bronk Ramsey C. 1998. Probability and dating. *Radiocarbon* 40(1):461–74.

Bronk Ramsey C. 1999. An introduction to the use of Bayesian statistics in the interpretation of radiocarbon dates. *Proceedings of the International Workshop on Frontiers in Accelerator Mass Spectrometry*. 6–8 Jan 1999. National Institute for Environmental Studies, Tsukuba. National Museum of Japanese History, Sakura, Japan. p 151–60.

Buck CE, Litton CD, Smith AFM. 1992. Calibration of radiocarbon results pertaining to related archaeological events. *Journal of Archaeological Science* 19:497–512.

Needham S, Bronk Ramsey C, Coombs D, Cartwright C, Pettitt PB. 1998. An independent chronology for British Bronze Age metalwork: the results of the Oxford Radiocarbon Accelerator Programme. *Archaeological Journal* 154:55–107.